

Rice University

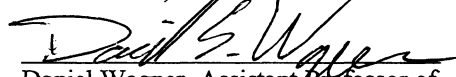
**A library approach to single site and combinatorial residue contributions
to dimerization of BNIP3-like transmembrane domains.**

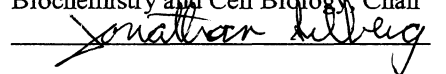
by

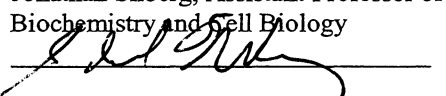
Christopher Paul Rodriguez

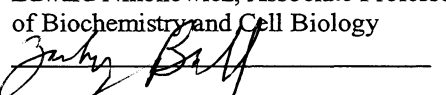
A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF THE
REQUIREMENTS FOR THE DEGREE
DOCTOR OF PHILOSOPHY

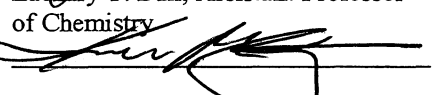
APPROVED, THESIS COMMITTEE:


Daniel Wagner, Assistant Professor of
Biochemistry and Cell Biology, Chair


Jonathan Silberg, Assistant Professor of
Biochemistry and Cell Biology


Edward Nikonowicz, Associate Professor
of Biochemistry and Cell Biology


Zachary T. Ball, Assistant Professor
of Chemistry


Kevin MacKenzie, Adjunct Professor
of Biochemistry and Cell Biology

HOUSTON, TEXAS

OCTOBER 2011

Abstract

A poly-leucine transmembrane domain library was randomized at positions corresponding to contact surfaces for a right-handed crossing of two helices to determine the significance of small residues, GxxxG motifs, and hydrogen bonding residues in driving helix-helix interactions within membranes. About 10000 sequences, which include the interfaces of tightly interacting biological transmembrane domains, were subjected to increasing selection strength in the membrane interaction assay TOXCAT and surviving clones were sequenced to identify single site and pairwise amino acid trends. Statistical analysis identified a central glycine to be essential to strong dimerization. The next strongest statistical preference was for a phenylalanine three positions before the key glycine. Secondary to these residues, polar histidine and asparagine residues are also favored in strongly dimerizing sequences, but not to the exclusion of hydrophobic leucine and isoleucine. The analysis identifies novel pairwise combinations that contribute to or are excluded from strong dimerization, the most striking of which is that the biologically important GxxxGxxxG/A pattern is under-represented in the most strongly associating BNIP3-like transmembrane dimers. The variety of residue combinations that support strong dimerization indicates that not only key 'motif' residues, but also the residues that flank them, are important for strong dimerization. Because favorable pairwise combinations of flanking residues occur between both proximal positions and residues separated by two or more turns of helix, the complexity of how sequence context influences motif-driven dimerization is very high.

Acknowledgements

I thank Laura Elisabeth Gracey for her unending support and help throughout my time at Rice and with many aspects of my life. I love you, and I feel as if many of the obstacles I faced the last ten years would have been insurmountable without you.

I want to thank my father whose dream of obtaining a doctorate became my own. Someday soon I hope we ride motorcycles together.

I thank my mother for showing me what it means to care about hard work. I never expected that but looking back I realize I should have. You do inspire me.

I thank my parents and sister for being patient with me while I found a balance between work and family.

I thank Dr. Kevin MacKenzie for being a leader and guiding me throughout my time at Rice. I will miss our scientific conversations, and I hope finds future success where he can educate and research where he will be happy.

I thank Dr. Todd Jaszewski for creating the RBS1 plasmid. I thank Kushagra Shrinath for subcloning the RBS1 plasmid and Sarah Lark for her contribution to preliminary work. Kush and Sarah were two of the best people I have had the pleasure to work with. I wish them success in the future as they finish their own educational path.

I thank my committee for understanding and supporting me, especially Dr. Jonathan Silberg who was pivotal to my success by championing my work for the last three years.

I want to thank the Shamoo, Silberg, and Wagner Labs for allowing access to equipment and advice. There were many times I needed an alternative perspective that Sol, Peter, and many others always had time for.

I thank the Houston Scooter Battalion and all the friends I have made there who became my support system. To my friend Jeffrey Powell - - thank you for the scientific talks. You helped me rediscover my passion for to abstract thinking. I dedicate this work to William Spoor, who left too soon.

Table of Contents

Chapter 1 Introduction and Outline	1
Chapter 2 Specificity and Stability of Protein-Protein	
Interactions within Membranes.....	4
2.1 Membrane proteins have unique sequence features	
related to their architecture	4
2.1.1 Architectural motifs in membrane proteins	4
2.1.2 Amino acid composition of transmembrane domains	6
2.1.3 Summary	8
2.2 TMDs make lateral associations in the membrane.....	9
2.2.1 Bitopic transmembrane domain dimers: A growing	
class of membrane proteins.....	9
2.2.2 Membrane hydrophobicity determines the rules of stability:	
α -helices as a transmembrane domain paradigm	10
2.3 Methods to identify and measure TM helix/helix	
interactions	11
2.3.1 Fluorescence Resonance Energy Transfer ties TMD	
association to the interaction of fluorophores.....	11
2.3.2 Sedimentation equilibrium ultracentrifugation investigates	
the number of TMD taking part in a complex	12
2.3.3 Thiol-disulfide equilibrium introduces a cysteine to track	
TMD association.....	13

2.3.4 <i>In vivo</i> TMD association assays.....	14
a. ToxR – an <i>in bacteria</i> prototypic TMD association assay.....	14
b. POSSYCCAT - POSitive Selection SYstem based on Chromosomally integrated CAT.....	15
c. GALLEX – an assay that incorporates the <i>lexA</i> DNA binding domain.....	16
d. TOXCAT – A self contained TMD association assay.....	17
e. A system-specific assay in mammalian cells based on the Platelet Derived Growth Factor Receptor β	18
2.3.5 Structural studies provide key insights but for few systems.....	19
2.4.4 Summary	20
Chapter 3 Forces and features driving association: from motifs to sequence context.....	21
3.1 Structural studies and spacing of key residues show that many biological TMD dimers exhibit a right handed crossing angle between helices	21
3.1.1 Glycophorin A	21
3.1.2 BNIP3	24
3.1.3 ErbB	26
3.2 Model systems and library approaches have identified roles for GxxxG motifs, strongly polar residues, and clusters of polar residues in TMD dimerization	28

3.2.1 A library approach reveals a role for GxxxG in strong TMD dimerization	28
3.2.2 Small polar residues can drive dimerization without glycine.....	29
3.2.3 A role for strongly polar side chains in TMD dimerization from design experiments.....	30
3.3 Variations on the GxxxG motif and understanding systems for which structures are not currently available	31
3.3.1 Biological examples where GxxxG contributes to dimerization	31
3.3.2 GxxxG motifs can operate in tandem	31
3.3.3 The GxxxG motif can be a red herring	32
3.4 The Problem of Sequence Context	33
Chapter 4 A library approach to understanding the sequence dependence of TMD dimerization	35
4.1 Advantages of library approaches.....	35
4.1.1 Libraries address large sequence spaces and robust statistical analysis is possible.....	35
4.1.2 The TOXCAT assay tests for TMD dimerization in cell membranes	36
4.1.3 <i>E. coli</i> is well suited to build and test libraries	37
4.2 Previous library studies have identified only the tightest associating TMDs and have used only very basic analysis methods.....	38

4.2.1 A TOXCAT library approach established the generality of the GxxxG dimerization motif.....	40
4.2.2 Dawson et al found a polar zipper motif in a library that lacks glycines	44
4.2.3 Herrmann et al find that histidine drives strong dimerization in specific cases.....	45
4.2.4 Summary of conclusions from past library studies.....	47
4.3 Over- and under-representation of residue pairs in biological TMDs	47
4.4 A new library to address open questions from previous structures, mutagenesis, and library selections of TMDs	49
4.4.1 Size requirements for a right handed TMD library from which clones that dimerize to different degrees will be studied in depth.....	50
4.4.2 Designing the interface and sequence context.....	51
4.4.3 Degeneracy of the genetic code and amino acid choice	53
4.4.4 Fusion protein expression level and antibiotic selection considerations	54
4.4.5 Statistical analysis.....	55
4.4.6 Summary/conclusions	56
Chapter 5 Plating Experiments and Single Site Analysis.....	58
5.1 Plating PGM and PGM-Low reveals a range of dimerization propensities	58

5.1.1 Surviving colonies drop off exponentially with increasing chloramphenicol	58
5.1.2 ToxR-TMD-MBP fusions are properly inserted in bacterial inner membranes	61
5.1.3 PGM and PGM-Low constructs display fast and slow growth on maltose media suggesting a qualitative difference in dimer stability	63
5.2 Sequencing reveals a mixture of sequence biases at different selection strengths	63
5.2.1 PGM single site residue trends reveal a positional hierarchy underlying self association of BNIP3-like TMDs	71
5.2.2 PGM-Low single residue trends	78
5.3 Comparison of PGM 400 and PGM-Low 200	82
5.3.1 Similarities/differences between winning sets.....	82
5.3.2 Selective pressure causes only slight differences between winning sets	83
5.3.3 Combining the libraries and analyzing with a “pooled winners” rationale	86
5.4 Comparing findings from PGM and PGM-Low to previous library studies	91
5.4.1 Statistical analysis of Russ et al’s leulib data reveals that only the strongest bias is in agreement with PGM and PGM-low libraries.....	91
5.4.2 Herrmann et al also identified a polar residue/glycine spacing in strong dimers	94

Chapter 6 Pairwise Analysis and Combinatorial Effects.....	95
6.1 Defining the hypergeometric calculation for top winners.....	95
6.1.1 Data collection and restructuring the hypergeometric to uncover pairwise interactions	95
6.1.2 Binning the effects into classes: over- and under-represented combinations	101
a. Over-represented trends	101
b. Under-represented trends	101
6.1.3 Interpretations of P-value magnitude	103
6.2 Comparing PGM/PGM-Low triplets to biological and library TMDs.....	103
6.3 Conclusion/summary	106
Chapter 7 Conclusions and Discussion	107
7.1 Framing the discussion: caveats and counter-arguments	107
7.2 Lessons learned about sequence space from PGM/PGM-Low analysis	108
7.3 Additional questions in this library	114
7.4 Low expression constructs are well suited for combinatorial studies.....	115
7.5 The present and future of library investigations of TMD interactions	117

Chapter 8 Methods	119
8.1 Library Design and Oligonucleotides.....	119
8.1.1 Design of transmembrane library sequences	119
8.1.2 PCR amplification of PGM Sequences.....	119
8.1.3 Purification, restriction digests, and ligations.....	120
8.1.4 Construction and transformation of PGM and PGM- low required 'high' competent cell lines	121
a. Preparation of highly competent DH5 α	121
b. Measuring competency	122
c. Building PGM (by multiple transformation approach)	122
Chapter 8.2 Plating Experiments	125
8.2.1 Preparation of electrocompetent NT326.....	125
8.2.2 Calibration of competency	125
8.2.3 Isolation of DNA.....	127
8.2.4 Bias Calculation.....	127
8.3 Maltose complementation	128
8.3.1 Preparation of minimal media cultures.....	128
8.3.2 Preparation of minimal media plates	128
References.....	129
Appendices.....	139

List of Tables

Table 4.1 The most tightly associating 24 sequences from Dawson et al. 2002

Table 4.2 Hermmann et al results

Table 5.1 Surviving Fraction

Table 5.2 Summary of Library Sequences

Table 5.3 Winning Sequences

Table 5.4 P-values for unselected residues arising from the intended library by chance.

Table 5.5 Single site residue biases in PGM under selection

Table 5.6 Single site residue biases in PGM-Low under selection

Table 5.7 Comparison of residue biases between top 1% of PGM and PGM-Low clones

Table 5.8 Top 1% sequences combined

Table 5.9 Single site biases for the combined PGM and PGM-Low libraries

Table 5.10 Leulib Analysis

Table 6.1 A. Pooled Unselected

B. Pooled Selected

Table 6.2 Pooled selected pairwise occurrence

Table 6.3 Statistical significance of pairwise biases

Table 6.4 Significant Motifs

Table 6.5 Significant Combinations

Table 8.1 Library Coverage

Table 8.2 Representative Plating Experiments

Appendix Table 1.1 A. PGM Total & Total Unique

B. PGM-Low Total &Total Unique

Appendix Table 1.2 Significance Analysis A.PGM

B.PGM-Low

List of Figures

Figure 3.1 GpA dimer structure

Figure 3.2 BNIP3 dimer

Figure 3.3 GxxxG depth

Figure 3.4 Depth of polar residues can affect dimerization

Figure 3.5 KcsA quaternary structure

Figure 4.1 Russ et al Library Design

Figure 4.2. Russ et al. results

Figure 4.3 Hermmann Library

Figure 4.4 PGM Library Design

Figure 5.1 Log-Linear decline

Figure 5.2 PGM/PGM-Low single site residue frequencies

Figure 5.3 A. PGM Over-represented single site residue biases

B. PGM Under-represented single site residue biases

Figure 5.4 A. PGM-Low Over-represented single site residue biases

B. PGM-Low Under-represented single site residue biases

Figure 5.5 PGM Library combination

Figure 5.6 Leulib single site residue frequencies with selected P-values

List of Abbreviations

BNIP3 – Bcl-2 nineteen kilodalton-interacting protein

BpE5 - Bovine papilloma virus E5 oncogene

CAM - chloramphenicol

CAT - chloramphenicol acetyl-transferase

FRET - Fluorescence Resonance Energy Transfer

GpA - glycophorin A

MBP - Maltose Binding Protein

POSSYCAT - POsitive Selection SYstem based on Chromosomally integrated CAT

TM - transmembrane

TMD - transmembrane domain

Chapter 1 Introduction and Outline

The most basic function of cellular membranes is to serve as boundaries, separating the contents of compartments within the cell from one another and from the immediate surroundings. Membranes also hold hundreds to thousands of membrane proteins that play critical roles in a myriad of physiological functions. Both the protein and the lipid composition of membranes vary with the function that each cell and organelle type performs, and this combination allows processes that occur at membranes to be highly specialized.

Transmembrane (TM) proteins are used by cells to sense the environment, to take up polar solutes, and to enable the cells in higher order organisms to distinguish self from non-self. Transmembrane proteins are pivotal to cell adhesion and motility, and the life cycles of viruses that lead to disease states are often mediated by both viral and endogenous transmembrane protein interactions.

TM proteins are involved in signal transduction. Ligand binding to extracellular domains induces conformational changes, which then lead to a change in oligomeric state or altered binding of effectors or adaptors in the cytosol. Changes in oligomeric state that alter the environment of the cytoplasmic domain of the TM protein can lead to intracellular protein-protein interactions and down stream effects such as auto-phosphorylation of tyrosine kinase domains. Much research has focused on the roles of extracellular and intracellular soluble domains of membrane proteins, especially when these domains are connected by a single TMD span. The membrane spans of such bitopic proteins are known to function as anchors, but in some instances they also make specific

protein-protein interactions that support function (MacKenzie and Fleming, 2008).

Evidence that the single membrane spanning region of certain bitopic TM proteins have an active role in oligomerization has been available for several decades (Furthmayr, 1977), but the difficulty of studying these interactions means that such roles continue to be discovered today. Given the functional importance of bringing TM proteins together, I expect that these numbers will continue to grow.

Several sequence motifs are currently thought to participate in driving association of single TM spans. These include the GxxxG motif, variants such as the glycine zipper and Gxxx(small) or (small)xxxG, and strongly polar residues (Senes et al., 2004). As described in Chapter 3, these motifs can support strong dimerization but do not do so in all instances. It is not currently possible to predict which TM span that contains a GxxxG motif or a strongly polar residue will self-associate strongly. The sequence context, or residues that flank the motif, must help determine this in a way that is not yet understood. In Chapter 4, I describe a designed library based on three naturally occurring transmembrane domain (TMD) dimerization interfaces that allows me to test the roles and relative importance of glycines, small residues, and strongly polar residues in dimerization, alone and in combination. I test the dimerization propensity of nearly 10^4 related TMD sequences using the TOXCAT assay, which confers antibiotic resistance to cells carrying constructs whose TMDs dimerize, enabling me to select for sequences that associate tightly. By analyzing sequences isolated at different stringencies, I have assessed the statistical significance of amino acid contributions at each of these positions, which are presented in Chapter 5. These tests have been carried out at two different protein expression levels, and colonies isolated at a range of antibiotic concentrations

allow me to assign statistical significance to individual positions and to pairs of amino acids. My data suggest that remote combinatorial effects (between residues more than two turns of helix apart) influence TMD association, and the significance of this is discussed in Chapter 7.

Chapter 2 Specificity and Stability of Protein-Protein Interactions within Membranes

2.1 Membrane proteins have unique sequence features related to their architecture

2.1.1 Architectural motifs in membrane proteins

Membrane proteins exploit either α -helices or β -strands to traverse the lipid bilayer; usually these motifs are mutually exclusive. Proteins that cross membranes using α -helices may have a single spanning region, such as in bitopic proteins. Other proteins use a series of α -helices, stitching through the membrane in a polytopic fashion. Membrane proteins that use β structure to span the bilayer do so by using many β -strands to build a cylindrical superstructure, referred to as a β -barrel.

Proteins that span membranes using a cylindrical arrangement of β -sheets form a polar tunnel across the membrane. In these β -barrels proteins, the protein backbone-backbone hydrogen bonds made within the membrane occur between adjacent strands, and the side-chains point either into the lumen of the barrel or outward from the surface of the barrel. Such β -barrels therefore contain three distinct amino acid classes: i) those that face the tunnel and are solvent exposed, ii) those that face lipid or membrane exposed, and iii) those that form the interface between β -sheets of different barrel monomers. The solvent exposed residues tend to be polar amino acids that can be stabilized throughout the pore of the molecule, while the residues that are at the hydrophobic face of the membrane are themselves usually hydrophobic. The interfacial residues between β -sheets are excluded from the solvent and lipid environments. These residues are made up of hydrogen bond donor-acceptor pairs (Wimley, 2003). Membrane

β -barrels can be identified from genomic sequences with modest reliability from the alternating pattern of hydrophobic and polar residues and the requirement for at least eight but as many as fourteen strands.

The vast majority of membrane proteins span the lipid bilayer using α -helical secondary structure, in which backbone hydrogen bonds are made between residues i and $i+4$. Helical transmembrane domains (TMDs) are primarily hydrophobic, consistent with their arrangement of side chains which all face lipids. However, polar residues do occur in TMDs, and certain arrangements of polar residues occur in nature more often than would be expected by chance (MacKenzie, 2006; Senes et al., 2000). Contacts between polar amino acids buried in a hydrophobic environment can be highly favorable (White and Wimley, 1999). For this reason polar residues are thought to play key roles in organizing and stabilizing interactions among helices of polytopic membrane proteins such as bacteriorhodopsin (Joh et al., 2008) or interactions between monomers of oligomeric single span TMDs such as BNIP3 (Sulistijo et al., 2003; Sulistijo and MacKenzie, 2006, 2009). The multiple TMDs of polytopic proteins are connected by cytoplasmic and ectoplasmic loops that can also help to organize the helical bundles that are buried in the membrane (Hirai et al., 2009; Tikhonova and Costanzi, 2009). Genome sequence analysis has revealed that membrane proteins may have as many as twenty hydrophobic TMDs (Arkin et al., 1997). Bitopic proteins, membrane proteins that traverse the membrane one time, make up about 40% of all membrane proteins. Such proteins can form oligomeric bundles of TMD helices, including homodimers (Lemmon et al., 1992a; MacKenzie et al., 1997; Sulistijo et al., 2003; Sulistijo and MacKenzie, 2009), heterodimers (Li et al., 2005), and higher order species (Arkin et al., 1994; Choma

et al., 2000; Oxenoid and Chou, 2005; Stouffer et al., 2008; Yin et al., 2007). In many cases these structures represent possible targets for the treatment of disease (Cady et al., 2010; Stouffer et al., 2008; Yin et al., 2007). How the sequences of TMDs help stabilize polytopic bundles and oligomeric complexes is a topic of considerable current interest (MacKenzie, 2006).

2.1.2 Amino acid composition of transmembrane domains

The strong hydrophobicity of helical TMDs allows these regions to be readily identified by the primary sequence (Engelman et al., 1986; Jayasinghe et al., 2001; Kyte and Doolittle, 1982; Snider et al., 2009). Hydrophobicity does not lead to self-insertion of a helix into the membrane. Instead, membrane insertion is mediated by additional cellular machinery (White and von Heijne, 2005). Hydrophobic residues are the most common constituents of TMDs, with leucine by far the most common of these, succeeded by isoleucine, valine, alanine, phenylalanine and glycine (Senes et al., 2000). Serine is roughly one third as common as leucine but is still more prevalent than the other amino acids with polar side chains. This may arise because serine can donate a hydrogen bond to the *i*-4 carbonyl oxygen in a helix (Gray and Matthews, 1984), which lowers the energy cost for submerging this side chain in an apolar bilayer. The bias against polar side chains continues as polarity and ionizability increase, concluding with arginine. Although strongly polar residues are found only rarely in membranes, none are absent from TMDs altogether (Senes et al., 2000). An apparent free energy scale that measures the tendency for the translocon to incorporate TMDs into the bilayer correlates well with general hydrophobicity (Hessa et al., 2005a), and even TMDs with several arginines can

be inserted across biological membranes if the other residues are hydrophobic enough (Hessa et al., 2005b).

Enrichment of TMDs with Leu, Ile, Val, Ala, and Phe allows for their stable accommodation into the bilayer as alpha helices. Backbone hydrogen bonds keep the helix from unfolding within the bilayer and the hydrophobic effect prevents the entire span from leaving the bilayer for the polar aqueous environment (MacKenzie, 2006; Popot and Engelman, 1990). Membrane thickness varies between cells with distinct functions since lipid composition is a feature directly tied to cell type. As a result the number of residues needed to span a membrane varies with their overall hydrophobicity and with the thickness of the specific bilayer being traversed (MacKenzie, 2006).

The membrane localized β -barrel, as described above in section 2.1.1, is a protein whose structure is made up of interlocking β sheet hairpins. These hairpins align to form a single aperture that allows polar solvated ligands to cross the membrane bilayer. β -barrels are essentially channels that border two different surroundings: the hydrophobic membrane and their own polar interior (Wimley, 2003). Even though β -barrels do not contain α -helices, the mechanism of adopting polar residues at one interface while maintaining apolar residues to separate a different boundary is a theme shared by both classes of molecules. Interestingly, a statistical tendency has been observed for aromatic residues to border the hydrophobic region of the bilayer or the interfacial region in both helical proteins and β barrels (Braun and von Heijne, 1999; Killian and von Heijne, 2000; Seshadri et al., 1998; Yau et al., 1998).

Although the interior hydrophobic environment of soluble proteins usually excludes highly polar residues, it has been found that even ionizable residues can be tolerated in otherwise hydrophobic cores (Stites et al., 1991). In a similar way, membrane proteins can contain ionizable residues in their TMDs. For helical TMDs, the energetic cost of transferring the side chain into the bilayer varies with the depth at which the residue resides (Hessa et al., 2007), perhaps because some residues can interact favorably with the negatively charged phosphates of lipid head groups. Even transferring a positively charged side chain to the center of a bilayer is possible if the cost is offset by flanking hydrophobic residues (Hessa et al., 2005a). The combination of interactions among polar and charge carrying residues within a hydrophobic environment are expected to provide the more important energetic terms to the complicated manner in which membrane protein structure is stabilized (White, 2005). However, amino acid identities at specific TMD depths that result in additional or detracted stability are beginning to be explored (Hessa et al., 2007; Senes et al., 2007). At this point the frequency of occurrence for residues at various positions along the TMD has been calculated. The functional reason for the appearance of an amino acid at a particular position is still not well understood.

2.1.3 Summary

Like soluble proteins, membrane localized proteins use α -helix and β -sheet secondary structures. The transmembrane domains of β -barrels are largely hydrophobic on one face and largely polar on the other, whereas α -helical proteins are typically made up of hydrophobic residues; however, polar residues appear occasionally in α -helical

TMDs and in the otherwise hydrophobic face of β sheet transmembrane domains. In both cases, polar residues are stabilized by interactions with other side-chains of opposite charge or polarity, although in some cases the stability contributions of polar interactions inferred from structure are not supported by direct experiment (Stanley and Fleming, 2007).

2.2 TMDs make lateral associations in the membrane

2.2.1 Bitopic transmembrane domain dimers: A growing class of membrane proteins

The number of discovered TMDs capable of associating in the lipid bilayer increases each year. TMD association is becoming widely accepted as a mechanism by which transmembrane proteins enact a function. Bitopic proteins known to interact through their TMDs include glycophorin A, the ErbB growth factor receptors, the pro-apoptotic protein BNIP3, the growth factor co-receptor syndecans, and several viral proteins (Dews and MacKenzie, 2007; Kochendoerfer et al., 1999; Laage et al., 2000; MacKenzie, 2006; Mendrola et al., 2002; Miyauchi et al., 2006). Dimeric TMDs of bitopic proteins are the smallest and simplest system in which to study helix-helix interactions. Single helices are not covalently bonded to another helix, and so their oligomers are not subject to complicating distortions of helix backbones exhibited by many polytopic membrane proteins (White, 2003).

2.2.2 Membrane hydrophobicity determines the rules of stability: α -helices as a transmembrane domain paradigm

The hydrophobic environment in lipid membranes constrains the type of protein sequences that lie within it, and in doing so influences the kind of contacts that are favored between membrane spans. Because of the complexity inherent in fulfilling these restraints and the difficulty in expressing and solving the structures of membrane proteins, TMD association as a whole is not well understood (MacKenzie, 2006; MacKenzie and Fleming, 2008). The structure of one of the best understood TMD dimers, glycophorin A, has been determined by NMR (MacKenzie et al., 1997) and the dependence of its association has been studied extensively using mutagenesis (Lemmon and Engelman, 1994; Lemmon et al., 1992a; Lemmon et al., 1992b). These studies showed that almost all polar substitutions are disruptive, whereas only ~40% of apolar substitutions adversely affected dimerization. The large number of mutations that were non-disruptive demonstrated the number of sequences that can be accommodated in α -helical TMD dimers. The structure of the glycophorin A dimer was used to interpret the effects of individual and multiple mutations on dimerization (MacKenzie and Engelman, 1998), and the inferred steric clashes that resulted from making residue substitutions within the solved structure correlated with known effects in biological and physical assays. The glycophorin A TMD was the first example where a dimer interface predicted based on biochemical association assay was validated by a structural study, and the success of similar approaches with BNIP3 (Lawrie et al., 2010; Sulistijo and MacKenzie, 2006, 2009) suggest that analysis of the sequence dependence of TMD dimerization can often lead to useful structural predictions.

2.3 Methods to identify and measure TM helix/helix interactions

In the following sections, I describe the current state of assays that can be used to measure association of single helical TMDs. Generally speaking, these methods represent a way for researchers to identify the features that contribute to transmembrane domain oligomerization by measuring an effect that is coupled to TMD association. Usually the effect is intrinsically linked to TMD association but some methods introduce a reporter as a means to extract data. The data collected using some of these methods can be used to determine the number of molecules that associate, here referred to as the ‘order’ of the oligomer.

2.3.1 Fluorescence Resonance Energy Transfer ties TMD association to the interaction of fluorophores

Fluorescence Resonance Energy Transfer (FRET) can be used to study protein-protein interactions *in vivo* or *in vitro* (Masi et al., 2010). In a FRET experiment, a fluorophore and quencher are conjugated to separate TMDs. After successful insertion into detergents or membranes, the spectroscopic profile is measured in real-time, and the amount of quenching is used to calculate the fraction of TMDs that form oligomers. The mathematical relationship between fluorophore and quencher as described by the data is also used to determine order (Adair and Engelman, 1994; Fisher et al., 1999, 2003; Li and Hristova, 2006; Li et al., 2006; Merzlyakov et al., 2007; Merzlyakov et al., 2006; You et al., 2005). The ability to measure FRET depends on the timescale of the interactions, but empirical data has shown that the binding is observable on the order of

minutes to hours. FRET can be performed in micelles or vesicles, provided that the detergent is fluorescently silent in the wave length region of the FRET species used. To guarantee high-quality data, this experimental approach must be repeated with varying conditions, *e.g.*, TMD concentrations and temperature. This type of exhaustive data collection is necessary in order to gather enough information to make a determination of oligomeric state. The oligomeric state is found by testing different models and determining which fits the experimental data best (Adair and Engelman, 1994; MacKenzie and Fleming, 2008).

The chief drawbacks of FRET are that microgram to milligram amounts of pure protein must be readily available, it must be possible to introduce fluorophores at appropriate positions, and it must be possible to insert labeled protein into membranes (or detergents) in a native state. For membrane proteins, this is not always plausible. The association of TMDs to be tested must also be reversible on the time scale of the experiment to be able to obtain thermodynamic properties (Fisher et al., 1999, 2003; Masi et al., 2010).

2.3.2 Sedimentation equilibrium ultracentrifugation investigates the number of TMD taking part in a complex

Sedimentation equilibrium using analytical ultracentrifugation is well-suited to study TMD oligomerization (Fleming et al., 1997). Sedimentation assays for membrane proteins strongly mirror experiments for soluble proteins, but the main divergence is the method of rendering the detergent buoyancy a non-factor. This is achieved by doping the aqueous solution with heavy water until the detergent buoyancy is matched; under these

conditions, the protein mass and oligomeric state contribute to the sedimentation properties but the detergent does not (Fleming, 2000, 2002). Sedimentation equilibrium allows a direct measurement of the mass of the protein or protein complexes in a variety of detergents, and in favorable conditions the thermodynamics of oligomerization can be extracted. These assays require milligram amounts of protein, to allow the testing of many buffer conditions during experimental optimization. Although each experiment gives rise to hundreds of data points it takes at least eight hours to reach equilibrium, and so several days of acquisition time are needed to obtain thermodynamic information because samples must be measured at multiple concentrations and at multiple speeds (Fleming, 2000, 2002). Finally, although adjusting buffer conditions can approach endogenous conditions, this method is innately an *in vitro* method limited to detergents and cannot be applied even to model membranes.

Although sedimentation assays have a potential to collect large data sets, this experimental approach requires the TMD association to be reversible on the time scale of the experiment. Like FRET, models for the oligomeric state can be tested against the data, and in favorable cases the results are conclusive (Fleming, 2000; MacKenzie and Fleming, 2008). In some instances, however, the results can indicate non-ideal behavior that cannot be modeled, and in other cases this method can fail to identify interactions that are detected in membranes (Kobus and Fleming, 2005; Stanley and Fleming, 2005).

2.3.3 Thiol-disulfide equilibrium introduces a cysteine to track TMD association

Thiol-disulfide exchange is also used to analyze TMD association by exploiting the formation of disulfides between monomers as a measure of the fraction dimer.

Varying the ratio of oxidized and reduced thiol in the samples and measuring the amount of cross-linked species under each condition enables the description of the oligomeric state and the free energy of oligomerization; this method has been successfully applied to dimers. However, it is necessary that the association of the TMDs to be studied is reversible on the time scale of the experiment (Cristian et al., 2003; MacKenzie and Fleming, 2008). Although potentially of great interest because it can be applied in either detergents or membranes, this method has not been as extensively employed as FRET or sedimentation approaches. Thiol-disulfide interchange requires the introduction of a cysteine residue, and requires that the amounts of monomeric and cross-linked species be quantitatively determined by some method, usually HPLC.

2.3.4 *In vivo* TMD association assays

Studies into the association of TMDs in living cells are primarily performed in *E. coli* membranes. These assays rely on the introduction of the TMD into a fusion protein, whose self-association then triggers the production of a readout protein. Care must be taken to control for overall production of fusion protein in cells to maintain reliable results. Limited progress has been made to develop assays in mammalian cells.

a. ToxR – an *in bacteria* prototypic TMD association assay

The discovery of a membrane-anchored dimeric transcription factor (Miller et al., 1987) led to the development of a number of TMD interaction assays, each with their advantages and disadvantages. The initial assay ToxR was built by encoding the DNA binding domain from the membrane sensing protein ToxR of *Vibrio cholera* in-frame with a transmembrane domain and the maltose binding protein (MBP). The fusion

protein is under basal regulation of endogenous ToxR promotor from *Vibrio cholera*. In this assay, TMD dimerization brings the ToxR domains together, allowing DNA binding and driving expression of the reporter gene *lacZ* from the *ctx* promoter. Increased self-association of the TMD leads to increased LacZ production, which can be measured colorimetrically or identified by screening for blue colonies on X-Gal plates (Langosch et al., 1996). The maltose binding protein domain of the fusion protein confers the ability to survive on maltose minimal media to *E. coli* strains lacking the *malE* gene, and this can serve as a validation that the fusion construct properly inserts into the *E. coli* inner membrane (Langosch et al., 1996). This assay has the powerful advantage of being performed in a bacterial membrane. However, its sensitivity is not as high as other assays based on ToxR (POSSYCCAT and TOXCAT assays). The genomic *ctx* promoter restricts the cell lines that can be used and restricts the study to association of single TMDs (Langosch et al., 1996).

b. POSSYCCAT – POSitive Selection SYstem based on Chromosomally integrated CAT

POSSYCCAT is a ToxR derivative that measures TMD association using the same coding region for the ToxR-TMD-MBP fusion protein but substituting the *araBad* promoter for the endogenous *toxR* promoter, so the fusion construct is induced specifically by arabinose and repressed by glucose. Varying the amount of arabinose allows the experimentalist to adjust the amount of fusion protein produced. Unlike other variable promoters, the temperature can be held constant. Constant temperature is paramount for studying TMD association since the thermodynamics of association could not be compared in experiments at different temperatures because this would introduce

multiple effects, e.g. increased bacteria growth rates, increased kinetic associations. In POSSYCCAT, the reporter gene whose expression is driven by the fusion protein is chloroamphenicol acetyl transferase (CAT) instead of lacZ and the reporter is chromosomally integrated (Gurezka and Langosch, 2001). For this reason, the method can only be applied to certain cell lines in the same manner as the ToxR system (Gurezka and Langosch, 2001).

c. GALLEX – an assay that incorporates the *lexA* DNA binding domain

GALLEX, a ‘ToxR like’ assay, was designed to measure homo- or heterodimeric interactions. To assess homodimerization, a single plasmid is used to express a TMD fused to MBP and the *lexA* wildtype DNA binding domain (*lexA*). The plasmid is transformed into a cell line containing a genomic copy of *lacZ* under control of the *lexA* promoter. Dimeric LexA represses expression from the *lexA* promoter (CTGTCTGT) (Dmitrova et al., 1998), so TMD dimerization results in repression of lacZ. GALLEX can also be used to test heteromeric association. However, this type of experiment requires multiple plasmids. The first plasmid encodes a TMD in the same manner as the homomeric version of this assay. The second plasmid in this case carries a LexA-TMD-MBP type fusion protein where the *lexA* DNA binding domain has been mutated to change its DNA sequence specificity (*lexA'*). In the heteromeric version of GALLEX, a modified cell line is used that contains a genomic mutant *lexA* operon (CTGTCCGT) that binds *lexA/lexA'* conjugates. This mutant *lexA* operon is not competent to bind homodimers of either *lexA* or *lexA'* and so homodimers are ignored in this assay. Heteromeric TMD dimers result in the suppression of lacZ expression. This system has a

tunable aspect built into it in the form of the plasmid origin of replication. GALLEX plasmids are available in pACYC184 or pBR322 based versions, which are low and high copy number plasmids, respectively (Schneider and Engelman, 2003).

Because GALLEX reads out dimerization as a repression effect, this assay is limited in the range of TMD association it can describe. The limitation is that all dimers that are strong enough to fully repress *lacZ* expression are forced to be categorized together. The different *ori* used could be used to globally reduce the degree of association, but the amount of plasmid in each experiment is difficult to determine. This makes interpretation of tight homomeric or heteromeric dimer interactions challenging.

d. TOXCAT – A self contained TMD association assay

A simple variation of the ToxR assay, TOXCAT, uses the same ToxR-TMD-MBP fusion as ToxR and POSSYCCAT assays but replaces the chromosomal *ctx:lacZ* reporter gene with chloroamphenicol acetyl transferase (CAT) *ctx:cat* located on the same plasmid as the fusion protein (Russ and Engelman, 1999). In this assay, the TMD of interest is cloned between the genes encoding the ToxR DNA binding domain and MBP. The plasmid can be transformed into any *E. coli* cell line, but using one that is *malE*- enables the MBP domain of the fusion protein to function as a control for membrane insertion. The level of CAT production can be measured by selection, for instance by plating cells on increasing concentrations of chloramphenicol (CAM). Alternatively cells may be screened by measuring CAT levels in cell extracts with a radioactive labeled assay (Russ and Engelman, 1999) or a linked spectroscopic assay

(Sulistijo et al., 2003). Higher CAT activity correlates with an increased degree of TMD association (Russ and Engelman, 1999).

One drawback of the TOXCAT assay is that it lacks a means to control fusion protein expression. This results in strongly associating TMDs having poor sensitivity to modestly disruptive mutations, as is the case of BNIP3 protein where point mutant TMDs known to be disruptive under other conditions give the same level TOXCAT signal as the wildtype version (Lawrie et al., 2010). A fusion protein population that is mostly dimeric at the constitutive levels of expression will exhibit minimal decreases in apparent dimerization with most mutations because the protein is well above the effective K_D . However, if the TMD in question does not drive TOXCAT to saturation, the method can be used to determine apparent free energy changes due to mutations (Duong et al., 2007). Because the CAT reporter gene allows for selection of tightly associating dimers, it is ideal for testing many TMDs at the same time. For this reason, this system has been used on several occasions with library approaches (Dawson et al., 2002; Russ and Engelman, 2000).

e. A system-specific assay in mammalian cells based on the Platelet Derived Growth Factor Receptor β

The PDGF receptor TMD assay is used to test association in mammalian cell membranes. This differs from the ToxR- and *lexA*- based assays which use *E. coli* membranes exclusively. The PDGF β receptor is a single span transmembrane receptor tyrosine kinase that becomes activated upon binding another transmembrane protein, the bovine papilloma virus E5 oncogene (BpE5) (Talbert-Slagle and DiMaio, 2009).

TMD/TMD interactions have been shown to drive PDGF/BpE5 binding, and PDGF TMD also self-associate functionally, with an effect on cell growth (Oates et al., 2010; Petti et al., 1997). Although this system has not been as popular as the ToxR type systems, it has been used to map important polar residues in transmembrane ligands of the PDGF receptor β (Freeman-Cook et al., 2004; Freeman-Cook et al., 2005) and to test the self-association of at least one heterologous TMD by substituting the wild type TMD of PDGF receptor β with that of p185neu. By expressing the PDGF receptor β mutant in PDGF receptor β null mouse cells, the effect on growth could be directly linked to the ability of the TMD to dimerize. The system has not seen wide application, as it is highly system-specific and requires a minimum of 6-8 weeks to produce a qualitative result (Petti et al., 1998).

2.3.5 Structural studies provide key insights but for few systems

Structural biology methods have had success on particular membrane proteins, but many membrane proteins are not well suited for these types of investigations. The first membrane protein structure appeared in the 1980s and since then classic structural biology approaches have had steadily increasing success, especially since the late 1990s (White, 2004). This has arisen because of improved methods for expression and purification of the target proteins and on better micro-focused synchrotron X-ray sources. X-ray structures of membrane proteins crystallized from detergents and from lipidic cubic phases have been reported, as have solution NMR structures in detergent micelles and solid state NMR studies in lipid bilayers. The rapidly increasing availability of structural details for membrane proteins places our understanding of these systems on

more solid footing and provides excellent starting points for structure-based investigations of membrane protein stability. This has been exemplified by a series of studies on the sequence dependence of the structure, stability, and folding kinetics of bacteriorhodopsin, an α -helical bundle protein (Allen et al., 2004; Faham et al., 2004; Joh et al., 2008; Sapra et al., 2008; Yohannan et al., 2004a; Yohannan et al., 2004b). However, many membrane proteins remain difficult to express with current methods making X-Ray and NMR studies unrealistic. At present, methods that complement structural biology must still be relied upon heavily to probe the nature of protein-protein interactions in membranes.

2.4.4 Summary

Sedimentation, FRET, and thiol-disulfide exchange are well-suited to thermodynamic studies in micelles. The *in vivo* ToxR-based assay gives good indications of TMD association in bacterial membranes and several choices exist that can be tailored to the experimental protocol. X-ray crystallography and NMR spectroscopy continue to be powerful research tools that can give definitive data but may take a long time to examine a single TMD interaction.

The application of appropriate ToxR type assays to libraries has led to a better understanding of sequence motifs that support TMD association. I have chosen to use a library approach with the TOXCAT assay, and I include a modification to control the overall production of protein chimeras. I will be exploring how altering TMD sequence results in gradations of TMD association strength while simultaneously determining how the amount of fusion protein produced affects association.

Chapter 3 Forces and features driving association: from motifs to sequence context

3.1 Structural studies and spacing of key residues show that many biological TMD dimers exhibit a right handed crossing angle between helices

3.1.1 Glycophorin A

The red blood cell protein glycophorin A (GpA) is anchored to the cell membrane by a single hydrophobic transmembrane domain. Besides acting as an anchor, the GpA TMD drives dimerization in detergents of both the intact protein (Bormann et al., 1989) and of a fusion protein containing a heterologous soluble protein and only the TMD (Lemmon et al., 1992a). Lemmon et al. used mutagenesis to identify the residues involved in TMD oligomerization by determining which changes had the greatest effect on detergent-resistant dimerization (Lemmon et al., 1992a; Lemmon et al., 1992b). After discounting the effects of strongly polar substitutions, which are uniformly disruptive, seven residues that correspond to one face of a helix (L₇₅IxxGVxxGVxxT₈₇) were the sites where mutations most influenced TMD dimerization (Lemmon et al., 1992a). Hydrophobic substitutions at no other positions significantly affected dimerization. The high degree of specificity is demonstrated by the effect of mutating the second glycine to alanine; this reasonably conservative substitution adds one methyl group but completely disrupts GpA TMD association. Because GpA is inserted in a single orientation (N terminus out) in red cell membranes, the GpA TMD dimer was expected to form a parallel, symmetric dimer. This idea gained support from two modeling studies using

different computational approaches. These studies analyzed mutagenesis-based data about detergent-resistant dimerization to test their models (Adams et al., 1996; Treutlein et al., 1992).

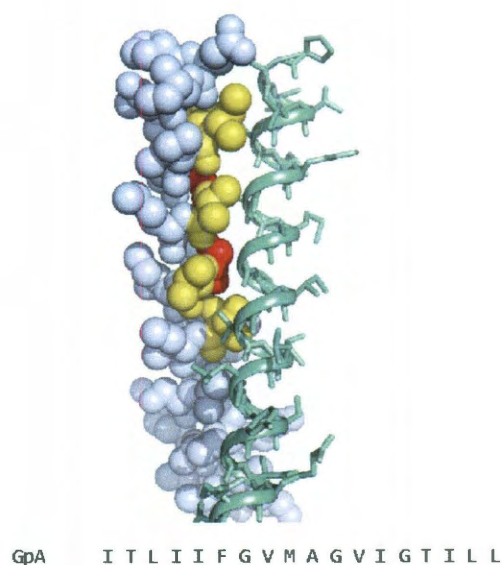
The ability of the seven residue GpA motif to drive dimerization of a poly-leucine TMD in detergents (Lemmon et al., 1994) further established the importance of the motif residues and the relative unimportance of the rest of the TMD. However, until the NMR structure of the detergent-solubilized GpA TMD dimer was determined (MacKenzie et al., 1997), it was not clear that the motif residues were in fact located at the dimer interface. The NMR structure of the dimer interface revealed that the GpA complex is closely packed, with the seven residues of the motif making intermolecular contacts in a ‘ridges-into-grooves’ packing. The glycines act as notches that allow the close approach of the two helices, and this close approach makes backbone-backbone intermonomer contacts (including $C\alpha-H \cdots O=C$ hydrogen bonds (Senes et al., 2001)) possible, see Figure 3.1. The remaining interfacial hydrophobic residues found adjacent to this point of closest approach fill in the widening backbone distances created by the 40 degree helix-helix crossing angle. The residues in the GpA dimer structure that make significant intermonomer contacts was found to be $L_{75}IxxGVxxGVxT_{87}$, the same motif identified by mutagenesis studies. The spacing of glycines found in GpA was subsequently shown to be highly over-represented in natural TMDs (Senes et al., 2000) and in a library study of TMD interactions (Russ and Engelman, 2000), leading to the idea that the $GxxxG$ motif may be a signature for dimerization.

The GpA dimer NMR structure was subsequently used to rationalize the effects of hundreds of point mutations (MacKenzie and Engelman, 1998). In this simplistic

approach, each mutation was built, one at a time, into the wildtype GpA TMD dimer NMR structure using side chain rotamers and without allowing the backbone to move. Each mutant was assigned integer scores for packing, clashes, and side chain entropy changes based on the modeled structure, and these scores were fit in a least-squares fashion against the mutagenesis phenotypes. Highly disruptive mutations correlated with significant steric clashes, and the resulting structure-based model could predict both stabilizing interactions and how some mutations that would otherwise disrupt dimerization could be accommodated by second mutations. The second mutation could do this by either being smaller, thus removing the possibility of steric clashes, or having an allowed rotamer to “swing” out of the way of bulky residues.

Figure 3.1 - GpA dimer structure

The NMR structure of the GpA TMD dimer (at right, N-terminus at top) reveals that the residues making intermonomer contacts are two key glycines separated by three residues (red). Five additional residues (in yellow) complete the interface (LlxxGVxxGVxxT). One monomer is shown as sticks for clarity.



The group of Karen Fleming has investigated the thermodynamics of how the GxxxG motif contributes to GpA dimerization *in vitro* using both single and double point mutations (Doura and Fleming, 2004; Doura et al., 2004). For single point mutations at

the interface that kept the GxxxG intact, the stability of the dimer could be decreased by as much as 4 kcal per mole from wildtype, showing that the GxxxG is not sufficient to drive dimerization of GpA (Doura et al., 2004). Most large-to-small substitutions are modestly disruptive, consistent with incremental losses in packing, and double alanine mutations are generally less disruptive than the sum of the two individual mutations, although two such combinations are much more disruptive than their sums (Doura and Fleming, 2004). These findings emphasize that the ways in which combinations of residues support dimerization can be complex. A TOXCAT analysis of the *in vivo* dimerization of wild type GpA and the single point mutants used in the studies described above shows that the biological assay TOXCAT obtains similar values for the apparent free energy changes associated with mutations (Duong et al., 2007), supporting the broad conclusions of the *in vitro* analyses.

3.1.2 BNIP3

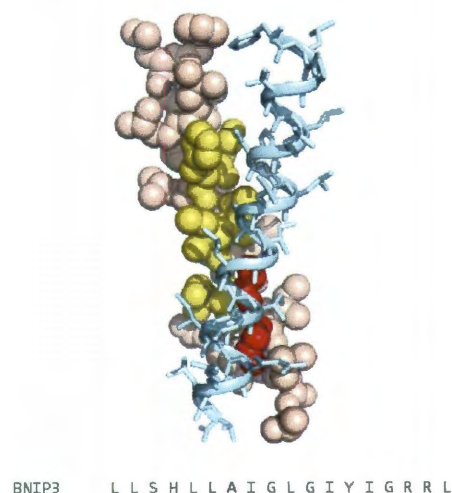
The presence of a GxxxG TMD signature in the mitochondrial pro-apoptotic BH3-only protein (BNIP3) led to investigation of the self-association of this sequence. Although the exact cascade of biological events involving BNIP3 are not entirely understood, the TMD of BNIP3 forms dimers in SDS detergent and in the biological TMD interaction assay TOXCAT (Sulistijo et al., 2003). Although the BNIP3 TMD contains a GxxxG motif, aligning it with the GxxxG of GpA suggests the interface AIxxGIxxGRxxT. From this sequence comparison it was not clear how the Arg would participate in the interface. An exhaustive mutagenesis study showed that the most important residues involved in detergent-resistant dimerization are S₁₇₂HxxAIxxGIxxG₁₈₄

(Sulistijo and MacKenzie, 2006). The SHxxAIxxGIxxG motif is shifted by one turn of helix relative to the expected GpA alignment (AIxxGIxxGRxxT). In the BNIP3 system, the intermolecular GxxxG is closer to the membrane boundary, which may alter the crossing angle. Regardless, this shift clearly introduces the polar residues histidine and serine into the interface. The BNIP3 TMD associates more than twice as tightly as GpA in TOXCAT, indicating that although both TMDs share the canonical GxxxG motif, other residues at the interface may alter the TMD behavior substantially.

The structure of the BNIP3 TMD dimer complex was determined by NMR spectroscopy, first with a sample that exhibited partial unfolding (Bocharov et al., 2007) and then with a well-ordered species (Sulistijo and MacKenzie, 2009). In this latter structure, motif residues (SHxxAIxxGIxxG) all make contacts across the dimer interface, see Figure 3.2. The polar residues serine and histidine were found to make intermonomer hydrogen bonds to each other. Extensive backbone-backbone contacts occur due to the close approach of helices, and at least one pair of symmetric non-canonical $\text{C}\alpha\text{-H} \cdots \text{O}=\text{C}$ hydrogen bonds was inferred to exist. The presence of these hydrogen bonding interactions helps to explain the increased TMD-TMD association strength.

Figure 3.2 - BNIP3 dimer

The NMR structure of the BNIP3 TMD dimer (at right, N-terminus at bottom). The interface is shown in red (GxxxG) and yellow (SHxxAIxxGIxxG). One monomer is shown as sticks for clarity.



A second mutagenesis study of BNIP3 TMD dimerization examined the sequence dependence of association in the TOXCAT assay and compared this to the previous findings in the SDS environment (Lawrie et al., 2010). Their results showed that the positional patterns of the effects of mutations in TOXCAT typically agree with the SDS-PAGE data, but the effects in TOXCAT are much less disruptive than they are in SDS-PAGE. This differs substantially from the analysis of GpA using TOXCAT (Duong et al., 2007), in which both stabilizing and destabilizing effects are seen in the membrane, in SDS, and in the ultracentrifuge. The authors suggested that the TOXCAT assay may be saturated by the strong self-association of BNIP3: the abundance of fusion protein being constitutively expressed would therefore cause the TOXCAT assay to be insensitive to many mutations effects (Lawrie et al., 2010). Truncating the BNIP3 TMD region by six residues on the N-terminal side, which leaves the motif intact, had little effect on wildtype dimerization but enhanced the disruptive effects of known mutations. Interestingly, the rank order of the effects of mutagenesis changes on TMD association remained the same relative to each other. This finding suggests that the absolute degree of interaction reported by TOXCAT can be fine tuned by altering the components of the system.

3.1.3 ErbB

The ErbB family of single pass transmembrane receptor tyrosine kinase proteins contains four members. The complete scheme of how signals are sensed from the outside environment by the ErbB extracellular domains and passed along the TMD to affect the cytoplasmic domain involves ligand binding, large scale conformational changes,

receptor homo- or hetero-dimerization, and interactions among the cytosolic tyrosine kinase domains (Lemmon, 2009). *In vivo* studies show that receptor oligomerization is the activating factor in this system, and deletion studies have provided strong evidence that the TMD is involved in dimer formation and signal transduction. Inspection reveals that the ErbB TMDs each contain two GxxxG motifs, denoted N- or C- for the proximity of the motif to the N- or C-terminal end of the TMD. A GALLEX study investigating the contributions of mutant N- and C- GxxxG motifs in bacterial membranes gave mixed results as to which motif was driving TMD homo-dimerization, suggesting that both motifs contribute to dimerization to some extent (Escher et al., 2009). Heteromeric interaction experiments from the same study gave similar results, suggesting that a complex hierarchy exist where specific N- or C- GxxxG-mediated TMD interactions could occur within or between ErbB members depending on the functional context (Gerber et al., 2004; Mendrola et al., 2002). A role for the ErbB TMDs in homo- and heteromeric association has also been proposed based on FRET experiments in detergent micelles (Duneau et al., 2007).

This type of GxxxG switching model gained momentum from an ErbB2 TMD dimer NMR structure (Bocharov et al., 2008). The assumed active state employed the C-terminal GxxxG motif at the TMD dimer interface. A computational modeling approach looking for stable TMD complexes found two minima, the C-terminal GxxxG interfacial state corresponding to the experimental structure and a state that places the N-terminal GxxxS motif at the interface, which the authors assign to an inactive state. Since the proposed states were closely matched in overall energy, previous researchers theorized that the whole protein complex would switch between interacting states upon ligand

binding (Fleishman et al., 2002). In this example, GxxxG switching could exploit two distinct sets of sequence contexts to modulate GxxxG-driven dimerization, dependent on which motif is buried at the interface in a given state. Unfortunately, the current state of the field lacks a code to predict which of the other interfacial residues may be having an effect on TMD association.

3.2 Model systems and library approaches have identified roles for GxxxG motifs, strongly polar residues, and clusters of polar residues in TMD dimerization

3.2.1 A library approach reveals a role for GxxxG in strong TMD dimerization

Using a selection method based on the biological TMD interaction assay TOXCAT, Engelman and colleagues showed that the GxxxG is over-represented in the strongest GpA-like TMD dimers

selected from a library of sequences with variable residues at the spacing of the GpA motif with intervening positions held constant (Russ and Engelman,

2000). The absolute position of the

GxxxG motif was found to vary

depending on if leucine or alanine was

selected as the host invariant residue, see

Figure 3.3. Although counter intuitive,

this suggests that sequence context effects are occurring from residues that are not in the

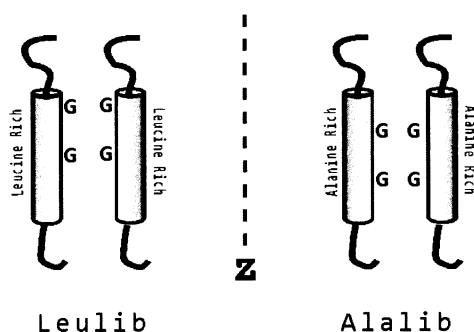


Figure 3.3 – GxxxG depth - The relative position of the GxxxG motif found in two libraries depends on the invariant residues. Such residues may affect TMD depth in the membrane or they may directly participate in forming interfaces; either could affect how the GxxxG motif drives TMD dimerization.

intended interface, so it is possible that for some sequences, the alanine or leucine host residues could be participating at the interface. In this approach, it was also found that the last residue of the GpA motif, threonine 87, occurs in 59% of sequences with a leucine background. There are two interpretations on the occurrence of GxxxG and threonine. Glycine and threonine could each be making independent single residue contributions to TMD association, or they could have combinatorial effects, as GxxxT, GxxxG, GxxxGxxxT, or GxxxXxxxT, but there is not enough data in the original study to carry out a detailed statistical analysis to establish which of these might be occurring. From the double mutant analysis of GpA (Doura and Fleming, 2004), the combinatorial effects seem likely to be important.

3.2.2 Small polar residues can drive dimerization without glycine

The dominant role for glycines revealed by the above experiments led the same group to try to identify residues other than glycine that are involved in TMD/TMD interactions. Using selection in TOXCAT of a library that excluded glycine residues from the TMD, they showed that combinations of serine and threonine residues could drive dimerization (Dawson et al., 2002). These selected motifs, SxxSSxxT and SxxxSSxxT, provide a more subtle layer to the motif/sequence context problem (Dawson et al., 2002). In several instances prolines occurred in combination with small polar residues (PSxxSSxxT and SPxxSSxxT), and these combinations were important for driving dimerization. *It is possible that many motifs exist that range in dimerization strength, each influenced by proximal and distal sequence context. At present, no structures are available for these types of polar zipper interactions.*

3.2.3 A role for strongly polar side chains in TMD dimerization from design experiments

Several lines of investigation have shown that aspartic acid, asparagine, glutamic acid or glutamine when introduced into TMD sequences that lack glycines or small polar side chains can result in strong dimerization and trimerization (Choma et al., 2000; Gratkowski et al., 2001; Zhou et al., 2000; Zhou et al., 2001). However, the exact location or depth within the membrane of the highly polar residues has a large effect on TMD-TMD equilibrium association, see Figure 3.4 (Lear et al., 2003), and the flanking sequence around a strongly polar residue can determine whether or not a particular strongly polar residue will enhance dimerization (Dawson et al., 2003). It is likely, but has not been conclusively established, that the strongly polar residues are at the oligomeric interfaces, and how sequence context influences the contribution of strongly polar residues to helix-helix interactions clearly needs to be understood to better anticipate TMD association strength.

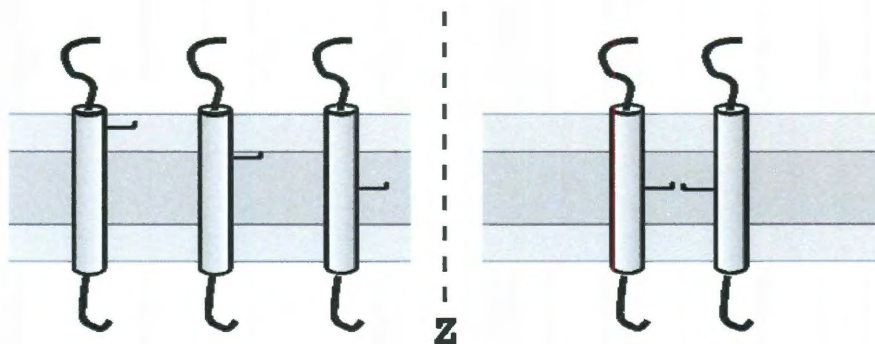


Figure 3.4 - Depth of polar residues can affect dimerization. Polar residues near the membrane/water interface (left panel, left and middle helices) do not drive dimerization as strongly as those buried in the core of the membrane (right panel). Such interactions are also influenced by the remaining TMD sequence context.

3.3 Variations on the GxxxG motif and understanding systems for which structures are not currently available

3.3.1 Biological examples where GxxxG contributes to dimerization

In section 3.1.1-3.1.2 I have already described instances where a single GxxxG motif influences natural TMDs to form strong dimers in membranes, for systems such as GpA, and BNIP3. As discussed in section 3.1.3, the ErbB tyrosine kinase receptor family TMDs also contain GxxxG motifs, but the presence of multiple motifs (or variants) within a single TMD may mean that these receptors switch between alternative GxxxG-mediated interfaces, where the balance between these states is controlled by the sequence of the TMDs and the interactions among the ecto- and cyto-plasmic domains. Other roles for GxxxG motifs have also been identified in biological systems, as outlined below.

3.3.2 GxxxG motifs can operate in tandem

In the cases of MscL, VacA, and MscS channel proteins, a tandem GxxxG motif (GxxxGxxxG) mediates formation of higher order TMD multimers (Kim et al., 2004). These channels are comprised of 5 to 7 TMDs where the tandem GxxxG motif takes part in driving a right handed type packing between the contributing TMDs. As observed with GxxxG-mediated dimerization, it appears that a small residue may in some cases substitute for glycine. KcsA uses four TMDs to form a channel and contains a GxxxAxxxT/A modified motif (Kim et al., 2005). Although exhaustive mutagenesis data is not available, some of the strongest evidence for the tandem motif effect comes from the high conservation between orthologs. Two statistical analyses have also reported that

tandem GxxxG motifs are over-represented in TMDs (Kim et al., 2005; Senes et al., 2000). It is not known if a tandem GxxxG motif can be taken as an indication of TMD association in general, or if a tandem GxxxG is an indicator of higher order TMD interactions. TMD sequences such as BNIP3, however, which contain not only GxxxG motifs but modified tandem motifs ($A_{176}xxxGxxxG$), show that some tandem motifs can form discrete dimers.

Figure 3.5 – KcsA quaternary structure

Four TMDs from separate monomers line the pore of the KcsA channel. These TMDs contain a motif (GxxxAxxxT/A) that allows close packing of the helices and stabilizes the tetramer. View of the pore from the cytoplasmic face (top) and from the plane of the membrane (bottom). From Kim et al., 2005.



3.3.3 The GxxxG motif can be a red herring

So far I have cited a number of instances in which TMD association is driven by GxxxG motifs. There are also, however, examples where the GxxxG motif does not drive TMD association. The SARS CoV S membrane protein, which is important for viral infection, forms trimers on gels through its TMD, which contains a GxxxG motif. However, this detergent-resistant interaction is retained upon mutagenesis of individual glycines to leucines (Corver et al., 2007). It has also been shown that mutants lacking the GxxxG are competent to cause infection of cells. The S protein TMD contains several

large polar residues not typical to TMDs known to be driven by GxxxG motifs that could be the driving force for S TMD association (Corver et al., 2007). One interesting possibility is that this TMD might adopt two different oligomeric structures, a trimer mediated by polar residues, and some other oligomer mediated by the GxxxG.

A predicted tyrosine kinase receptor protein, colon carcinoma kinase 4 (Cck4), contains a TMD GxxxG motif. A study that tested the ability of this TMD to form dimers using an ultracentrifugation assay showed no evidence for preferential dimer formation (Kobus and Fleming, 2005), with the small fraction dimers observed being explained by non specific association by overcrowding the detergent. In this case, the ineffectiveness of the GxxxG to cause dimer formation is not well understood (Kobus and Fleming, 2005), although it is possible that this sequence is tailored to form heterodimers but not homodimers. As pointed out in the description of GpA dimerization in section 3.1.1, some point mutations that leave the GpA GxxxG intact essentially abolish dimerization, so it is not unreasonable that some occurrences of GxxxG motifs would be unrelated to TMD self-association.

3.4 The problem of sequence context

In this chapter I have described several known TMD oligomer structures and our current understanding of how the GxxxG motif and its sequence context contribute to TMD association. The current understanding in the field of TMD association is limited by the few available detailed TMD oligomer structures: there are still more putative motifs than known structures, and the degree to which lessons learned in one system are

general and can be transferred to other systems is not clear. The importance of not only motifs but also sequence context that has been revealed by the most detailed mutational analyses to date (GpA, BNIP3) suggests that it is not possible to make useful predictions about the self-association of any TMD based on a motif alone. For any biological TMD that has been analyzed, the motifs within that TMD have been within a distinct sequence context. Predicting whether some other TMD will dimerize is difficult since it is likely that small sequence changes anywhere in the TMD could alter the interface enough to affect the forces that drive association. Currently two global questions remain unanswered: Which residues when placed at specific positions contribute to TMD association? What pairs of amino acids have synergistic effects on TMD association? To better understand TMD association, I have undertaken library studies to supplement the known mutagenesis and structural data and to make an assessment of sequence elements on TMD association that relies less exclusively on 'motifs' and considers 'sequence context' and 'pairwise contributions' more explicitly.

Chapter 4 A library approach to understanding the sequence dependence of TMD dimerization

This thesis presents the findings of a library-based, selection-driven approach to understanding the contributions of glycines, small polar residues, and strongly polar residues to TMD dimerization. The experiments were designed to test two aspects of the current state of knowledge in the field that I described in Chapter 3: the importance of particular motif residues, and the importance of flanking residues on the same face of the helix. My library approach allows me to assess the importance of contributions from a large number of positions on one face of an α -helix, and by examining the roles for residues at these positions at different TMD interaction strengths I will show for the first time how these contributions vary with the tightness of the interaction. Here I describe the strengths and advantages of library approaches, identify aspects of previous library methods that can be improved upon, and present my rationale for the design of the library and the experiments used to probe it. The experimental data themselves are presented in Chapters 5 and 6.

4.1 Advantages of library approaches

4.1.1 Libraries address large sequence spaces and robust statistical analysis is possible

The main rationale for using libraries to study protein folding, *in vitro* evolution of enzyme activity, or the self association of TMDs is that by building and testing large sets of sequences, the researcher can identify those sequences with particular properties

and identify the principles or elements that confer the properties on these library members but not others. The large amounts of sequence data that can be generated by such approaches can also allow the researcher to exploit powerful statistical analysis tools to calculate the significance (P-values) associated with these findings. In the work described in this thesis, I use the hypergeometric function to calculate the significance of finding particular residues (or pairs of residues) in populations of sequences from unbiased and selected data sets as described further in section 4.3.5. Larger sample sizes improve the reliability of the P-value calculation; and for the work described in chapters 5 and 6, I consider P-values < 0.05 based on about 75 sequences to be biased, and P-values < 0.001 to be highly biased; P-values less than 10^{-10} are calculated for many of the single-site analyses in Chapter 5. Robust statistical approaches to analyzing the data allow me to clearly identify the relative significance of trends in the data.

4.1.2 The TOXCAT assay tests for TMD dimerization in cell membranes

I built libraries in the TOXCAT biological reporter system (Russ and Engelman, 1999) because it readily allows me to select for sequences with variable degrees of TMD self-association (Dawson et al., 2002; Russ and Engelman, 2000). In the TOXCAT system, increased TMD-driven dimerization of the fusion protein causes increased expression of the reporter gene chloramphenicol acetyl-transferase (CAT), which results in increased resistance to the selection drug chloramphenicol. A major methodological improvement in my library studies compared to those undertaken previously is to sequence clones from several different levels of selection stringency, which allows me to track changes in the significance of particular residues as the association strength

increases. In addition, the maltose complementation assay built into TOXCAT enables me to determine what fraction of sequences in my library are made and inserted into the membrane, and to confirm that every sequenced clone used in the statistical analysis is correctly inserted in the membrane.

4.1.3 *E. coli* is well suited to build and test libraries

The bacterium *E. coli* is chosen for these experiments primarily because it allows us to use the TOXCAT biological reporter system (Russ and Engelman, 1999), which has been used for several library experiments previously (Dawson et al., 2002; Russ and Engelman, 2000). Working in *E. coli* also means that preparation of highly competent cells can be done in house, which allows us to maintain a low cost while ensuring the quality of the cells used in library experiments. High quality and efficiency in competent cells is critical to building a library, where transformations of ligated vectors and inserts require high efficiency and dependability, and also to assay experiments, when we transform the constructed library in its entirety into cells where antibiotic resistance (TOXCAT reporter gene expression) can be assessed. An *E. coli* cell doubles on the order of 20-40 minutes depending on the strain, temperature, and media chosen. The fast reproduction of bacteria enables us to perform cloning and selection experiments over several days.

It would also be valuable to study TMD association in a mammalian cell line, since the lipid composition and bilayer thickness membranes differs considerably from source to source, and even from organelle to organelle. However, a robust mammalian assay for membrane protein-protein interactions is not currently available. Although the

absolute degree of TMD self-association in *E. coli* membranes and in detergents can be quite different, good agreement in the rank ordering of the effects of point mutations on dimerization in very different lipidic environments for the GpA TMD (from *E. coli* membranes, SDS-PAGE, and two detergents used for ultra-centrifugation (Duong et al., 2007)) and for the BNIP3 TMD (in *E. coli* membranes and SDS-PAGE (Lawrie et al., 2010)) suggests that many conclusions about the sequence-dependence of TMD-TMD interactions in *E. coli* membranes can be taken as general rules of TMD association. At the same time, it must be acknowledged that the overall strength of TMD-TMD interactions are probably altered by lipid head group composition and bilayer thickness, and that there is a potential for any effects we see here to be specific to the particular *E. coli* attributes inherent to this experimental design.

4.2 Previous library studies have identified only the tightest associating TMDs and have used only very basic analysis methods

Previous attempts to analyze sequence effects on TMDs using libraries have identified the most strongly associating dimer (Dawson et al., 2002; Herrmann et al., 2009b; Russ and Engelman, 2000). This has been achieved in the TOXCAT system by starting with a very large library and sequencing only the tiny fraction that survives at very high level of the drug chloramphenicol (CAM). The CAM selection scheme allows the experimentalist to arbitrarily raise the stringency so that only clones with the most tightly associating TMDs, which will produce the most chloramphenicol acetyl-transferase (CAT), survive and are sequenced. The studies that I describe in this section outline how such approaches have allowed the field to gain an understanding of trends for

single residues and patterns or motifs that are present in very tight dimers. The results are usually examined by looking for common elements among the most strongly associating library members; this leads to the recognition of motifs but tends to blur or paper over the effects of sequence context. In the library approach that I use for my experiments, I not only identify motifs, but by sequencing at a series of different selection strengths, obtain information about how important these motifs are at different TMD association strengths. I also introduce approaches to extract pairwise correlations that indicate how sequence context underlies motif-driven TMD self association. Neither varied stringency nor this type of statistical analysis has been previously used to mine TMD sequences obtained from library selection schemes.

There are considerations that need to be examined due to the nature of library approaches. Deciding on the size of a library is greatly influenced by the type of information being sought. Previous library experiments carried out in bacterial membranes have used high throughput approaches, designing libraries that contain as many sequences as possible so that the selection protocol used for the particular ToxR-based TMD association assay can find the best possible ‘winners’ (Dawson et al., 2002; Russ and Engelman, 2000). In order to keep the possible sequences to a manageable number, only residues at the spacing of the expected interface (one side of a helix) are allowed to vary in the library. Non-interfacial residues are chosen to be a hydrophobic amino acid, typically leucine, which is well tolerated in TMDs. Choosing leucine as the flanking residue establishes a specific background sequence context that is reasonable for my experiments and makes my work directly comparable with several other studies. Because I intend to sequence not only at the highest possible selection stringency, but

also at lower stringencies, I have chosen to build a small library so that sequencing a reasonable number of clones (about 70) represents a significant fraction of the clones isolated at any particular condition.

4.2.1 A TOXCAT library approach established the generality of the GxxxG dimerization motif

Given that the seven GpA interfacial residues identified by mutagenesis (Lemmon et al., 1992b) and validated by the GpA TMD dimer NMR structure (MacKenzie et al., 1997) can drive strong dimerization when grafted into a poly-leucine TMD (Lemmon et al., 1994), an obvious question to ask is: “What other combinations of residues at this spacing can drive strong association of TMDs?” To answer this question, Engelman and colleagues used a library-based selection scheme to identify the sequence motifs that could drive dimerization of a right-handed type TMD dimer (Russ and Engelman, 2000). This study took the spacing of the GpA motif residues (LI..GV..GV..T) and substituted the interfacial residues with degenerate choices (XX..XX..XX..X) and the flanking residues with either polyleucine or polyalanine, synthesizing the entire TMD as oligonucleotides to be cloned into TOXCAT. Each degenerate position was allowed to sample one of 9 amino acid options: G, A, V, L, I, S, T, P, or R. The rationale for these residue choices was that building the library and selecting for tightly associating sequences would allow the researchers to determine how combinations of small residues (G and A), large residues (V, L, I), small hydrogen bonding residues (S, T), and backbone-altering residues (P) could contribute novel motifs to TMD association (Russ and Engelman, 1999). Given the genetic code, the library also needed to include the

	12	34	56	7					
Alalib	AS	xxx	AA	xxx	AA	xxx	AA	x	AILI
Leulib	AS	xxx	LL	xxx	LL	xxx	LL	x	LILI

x=G, A, V, L, I, S, T, P, R

Figure 4.1 – Russ et al Library Design - Leulib and Alalib were designed with randomized positions spaced according to the GpA interface and flanked by invariant positions. Proposed interfacial positions (numbered) are allowed to sample residues that appear frequently in TMDs. The Alalib and Leulib libraries encode the same interfacial possibilities but use alternate background residues (flanking x). These background residues, which are expected to be largely excluded from the dimer interface, are alanine (alalib) and leucine (leulib). Each library encodes 4.8×10^6 unique sequences. Adapted from Russ et al., 2000.

large charged residue R. Note that the motif sequence of GpA itself can be found in this library, ensuring that at least one very tightly interacting dimer will be encoded.

Including more strongly polar residues in the library would have increased the library size dramatically and might have generated sequences that would not partition into membranes. Two libraries were built composed of $\sim 10^7$ unique sequences each and then ligated into the TOXCAT assay plasmid. These libraries, designated leulib and alalib, differed by the choice of background residue (leucine or alanine) used at the flanking or host positions (Russ and Engelman, 1999).

The researchers independently transformed into *E. coli* and spread on CAM plates of increasing concentration in 50 $\mu\text{g/ml}$ steps. Surviving colonies followed a log-linear drop off, with ‘winners’ defined at 350 $\mu\text{g/ml}$ and 400 $\mu\text{g/ml}$ for leulib and alalib, respectively. At these antibiotic levels, only the top 0.001 % most strongly self-associating sequences appear (Russ and Engelman, 1999). The limited number of clones mined from this study was insufficient to make a statistical analysis.

For leulib, sequencing 47 clones revealed that positions 3 and 5 showed a high preference for glycine, and if glycine was absent serine was present, see Figure 4.2. The GxxxG motif was found in 96% of collected winning sequences in this library. Taking the serines in account, SxxxS or SxxxxS was possible in this library. Where as in another study (see below) these motifs were considered to drive strong association, they occur very infrequently in the top selected sequences of leulib even though the library encodes a great many of these motifs (Dawson et al., 2002; Russ and Engelman, 2000). The serine motifs’ relative strengths are unclear since the glycines dominated most sequences. This demonstrates the importance of sequence context in finding motifs. The GxxxG present

in leulib aligns with the GxxxG motif in the GpA dimerization motif, and the most common amino acid at position 7 was threonine, also the wildtype GpA residue. At the remaining positions, wildtype GpA residues were neither clearly excluded nor overwhelmingly over-represented because no residues dominated the findings. Several positions excluded residues entirely: position 2 (threonine), and position 4 (proline). Unlike positions 3, 5, and 7 where GpA residues dominated, there is not an apparent reason for the exclusion of other residues.

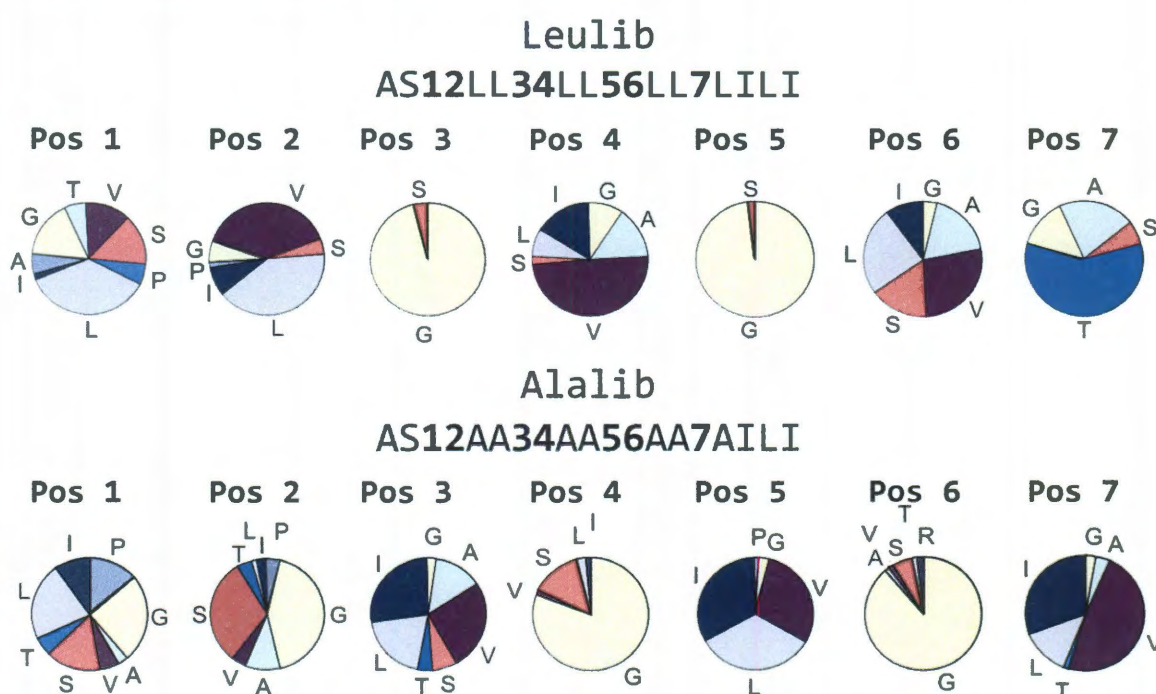


Figure 4.2. – Russ et al. results - Single site residue biases in tightly dimerizing sequences from the Leulib and Alalib libraries of Russ et al, 2000. Clones selected at high stringency were sequenced, and the frequency of residues identified at each position is proportional to the area in the pie chart. Some positions show many residues in similar proportions (Leulib position 1, Alalib positions 1 or 3), others show biases towards one class of residue (Leulib position 7, small; Alalib position 5, hydrophobic), and others are strongly biased towards one residue (Leulib 5, glycine). Adapted from Russ et al., 2000.

The alalib library results were based on 71 sequences, and 86% of these sequences contained a GxxxG motif, showing that this motif is important in the context

of an alternate host residue. However, these glycine pairs usually occurred at positions 4 and 6, shifted by one residue from the leulib motif positions. In alalib, serine was common to sequences that did not display a GxxxG, in much the same way as leulib. Similar to leulib, several positions in the most tightly associating TMDs exclude certain residues. These exclusions take place at position 2 (isoleucine), position 3 (glycine), position 5 (glycine, alanine, serine, threonine), and position 7 (glycine, alanine, proline).

4.2.2 Dawson et al found a polar zipper motif in a library that lacks glycines

Given the dominant importance of glycine to TMD dimerization in the study of Russ et al (Russ and Engelman, 2000), that group undertook another TOXCAT study to identify important interfacial residues in a library that lacked glycines (Dawson et al., 2002). In this library, the residues allowed at variable positions (A, T, S, F, V, L, I, P) result in 2×10^6 unique sequences in a leucine background. Performing selection experiments in a similar fashion as Russ et al (Russ and Engelman, 2000), the most tightly associating 24 sequences were analyzed (see Table 4.1) and two motifs, SxxSSxxT and SxxxSSxxT, were inferred from these data. Mutagenesis experiments were carried out on two sequences, isolate 3 (LALLSSLLSSLT) and isolate 8 (LSLLSPLLSSLT), to determine which of these residues are critical to dimerization. For isolate 3, single point mutations of most serines to alanine disrupted dimerization, as did mutating the threonine to either valine or serine. For this isolate, each of the ‘polar zipper motif’ residues seems to be important for dimerization. In contrast, single point mutations of serine to alanine did not disrupt dimerization of isolate 8. This suggests that

isolates 3 and 8 use different mechanisms for association, so it is difficult to infer a generalized motif for dimerization from these examples.

Table 4.1 The most tightly associating 24 sequences from Dawson et al. 2002.

ISLLSSLLSSLLTL	FILLPSLLSSLLTL	TILLTALLTFLLTL
PALLSSLLSSLLTL	VALLPSLLSSLLTL	LALLFPLLPVLLTL
LALLSSLLSSLLTL	AALLPSLLSSLLTL	LPLLFPLLVILLAL
LALLSSLLSSLLTL	FSLLAPLLSSLLTL	LPLLFPLLVFLLAL
VILLTSLLSSLLTL	TTLLAPLLSSLLTL	VILLAVLLVFLLLL
PSLLSPLLSSLLTL	PSLLAPLLSSLLTL	PSLLSPLLASLLTL
FSLLSPLLSSLLTL	SPLLPAALLSSLLTL	FALLPSLLSSLLTL
LALLSPLLSSLLTL	LVLLSALLSSLLTL	

4.2.3 Herrmann et al find that histidine drives strong dimerization in specific cases

A study that used a low expression version of the ToxR assay investigated TMD association by randomizing eight residues on one face of a helix by allowing all 20 amino acids simultaneously and including stop codons (Herrmann et al., 2009b), see Figure 4.3. This library size was theoretically 2.5×10^{10} possible sequences although this group found only 1.5×10^5 clones that inserted properly into the membrane. The number of unique sequences is lower than this since strong biases were found due to the PCR randomization technique used. For clones with tightly associating TMDs, histidine was over-represented at position 3, and replacement with leucine at this position resulted in decreased association. Presumably the polar or hydrogen bonding properties of this residue were essential to dimerization at this position.

Of the strongly dimeric sequences, two ‘exemplary’ TMDs LS46 and LS52 were further examined in a follow up study (Herrmann et al., 2009a). Both sequences gave

TOXCAT dimerization signals twice that of glycophorin A. For the first example, LS46, the positions and identities of ionizable residues that contribute to dimerization were important: when these polar residues were swapped or altered, dimerization was

1 23 45 67 8
Herrmann ASxLLxxLxxLLxxLLxGILI
Library

x=All residues

Figure 4.3 – Herrmann Library - Designed to contain 2.5×10^{10} unique sequences, the authors cloned 1.5×10^5 TMDs that could properly insert in to the membrane. The spacing of variable residues was chosen to match contacts for a left-handed crossing of helices; note that there is only one ‘background’ residue between positions 3 and 4. In spite of large bias due to the construction technique, and the likelihood that the designed sequence space was not fully sampled, screening for tight dimerization found a propensity for histidine at position 3. Adapted from Herrman et al., 2009.

reduced, refer to Table 4.2. In the second example, LS52, some replacements were tolerated, but swapping polar residues reduced dimerization.

Table 4.2 - Herrmann et al results - Selected mutations and approximate ToxR signal normalized to parental sequences. Adapted from Herrmann et al., 2009.

<u>TMD</u>	<u>β-gal Signal</u>
LS46	1.00
LS46-D5E	0.62
LS46-D5R	0.30
LS46-R6D	0.40
LS46-R6K	0.80
LS46-D5R/R6D	0.45
LS52	1.00
LS52-R6K	1.05
LS52-E8D	0.87
LS52-R6E/E8R	0.65

4.2.4 Summary of conclusions from past library studies

Previous library studies have been used to discover motifs that contribute to strong TMD dimerization, but these motifs may represent only a subset of physiologically relevant TMDs that interact. Many motifs identified by different studies contain small residues glycine or serine but some contain larger polar residues such as histidine. Because the selection schemes employed in these studies are different, it is not possible to rank the interaction strengths of these TMDs or their motifs relative to one another. Some authors have attempted to determine if particular sequence elements contribute synergistically to association, but only by making point mutations in a few selected clones rather than by mining their libraries as a whole for evidence across all clones. In my library approach, I look at lower strength TMD dimers in a novel library to determine the biases in selected sequences that occur as the degree of TMD association increases.

4.3 Over- and under-representation of residue pairs in biological TMDs

The sequencing and annotation of online genomic databases has made a large set of TMD sequence information available. By treating this data as a library set, many correlations can be separated from random expectations. A study into over- and under-represented pairwise residue correlations in TMDs was carried out using collected TMDs that were contained in the Swiss-Prot database (Senes et al., 2000). After removing repetitive and homologous entries from a set of inferred bitopic and polytopic proteins, the resulting set, TM-Stat, contained about 13,600 TMDs from eukaryotic, bacterial, archaeal, and viral sources. These sequences were analyzed for biases in the spacing of

pairs of residues compared to the random distribution of spacing expected based on composition.

The general composition of these TM-Stat sequences provides an indication that a preference hierarchy of residue pairings occurs in TMDs (Senes et al., 2000). It is generally accepted that helical TMDs are primarily hydrophobic, and TM-Stat supports this belief. In addition, TM-Stat contains a small percentage of polar residues. Statistical analysis of the residue spacings discovered strong biases, the most significant of which is over-representation of the GxxxG motif. If this sequence occurred only at random, about 1250 GxxxG occurrences would be expected; however, the motif appeared about 1650 times. The probability of this many pairs occurring at random is extremely small ($P\text{-value}=6 \times 10^{-34}$), indicating that the GxxxG is statistically over-represented in the general set of biological TMDs. Analyzing deeper, researchers found that glycine spaced by 3 residues to glycine was the only spacing that fell outside expected distribution for glycine to glycine combination instances – that is, GxxG and all other pairs except GxxxG appear at frequencies consistent with random expectation. Many pairwise combinations of other residues are over- or under-represented with very high significance ($P\text{-value} < 1 \times 10^{-10}$). The pairs GxxxA and AxxxG were both found to occur more than expected, suggesting that, like GxxxG, these pairs may support helix-helix interactions, but without any experimental measurements for the helices in this large data set, the reason for any of the over- or under-representations is unknown. Because GxxxG is known to contribute to TMD association in a few other systems, it is reasonable to assume that dimerization is the justification behind GxxxG over-representation in TMDs, although we cannot infer how tightly a given GxxxG motif in the TM-Stat data set might drive TMD dimerization.

It is important to recognize that this library was collated from natural sequences without any regard to selective pressure for dimerization. The other over- and under-represented motifs are suspected to have functional rationales behind their prevalence, but the specific reasons behind any particular motif being over-represented are unknown at this time. This approach demonstrated that using a very large set of sequences and robust statistical analysis allows very strong P-value significances to be discovered, but at the same time this data set does not allow these significances to be directly correlated to self-association or function. Although I expect that TM-Stat contains many TMDs from membrane proteins of known structures, the research group responsible for this study did not include this in their analysis, probably because the known TMD structures category would make up a very small percentage of the TM-Stat data set.

4.4 A new library to address open questions from previous structures, mutagenesis, and library selections of TMDs

Previous mutagenesis, structural and library studies have identified sequence motifs potentially involved in TMD dimerization, but there is no unifying description of how these indicators (polar pairs, GxxxG, tandem motifs) combine to produce TMD dimers nor is there a basic understanding of how to rank the strength of each indicator relative to one another. Previous library approaches picked only a small top tier of strongly dimeric sequences for analysis from a very large total number of sequences, which allowed them to identify ‘winners’ but which hid the relative contributions of different sequence elements to weaker levels of dimerization. In my work I expand on previous library approaches to describe how the sequence landscape changes with

increased stringencies on TMD associations. I have evaluated the strongest of these sequences and have identified several pairwise correlations capable of enhancing and disrupting TMD interactions.

4.4.1 Size requirements for a right handed TMD library from which clones that dimerize to different degrees will be studied in depth

In this work, I have tailored a library approach to study the combinatorial effects of the sequence dependence of TMD interactions by minimizing the complications inherent in large sequence pools. The typical TMD contains 15-18 amino acids, any of which could correspond to one of twenty amino acids. Given that each position can take a natural amino acid, we have the dilemma of assaying at least 20^{15} different sequences if one considers the full sequence space of a TM helix. Disregarding the complication of assaying such large numbers of sequences, simply generating 20^{15} TMD sequences by current DNA manipulations is an unobtainable goal, so the size of the library under consideration must be decreased.

If we consider only the interfacial residues as degenerate positions in a TM helix, then with seven residues contacting one another at a GpA-like interface, we arrive at about 20^7 unique sequences. This is an achievable sequence diversity to build in a library using PCR. However, the goal of my research is to understand how residues and combinations of residues contribute to helix-helix interactions as the strength of the association increases. This puts additional constraint on the library size in order to avoid the potential need to obtain thousands of sequences at varying selection levels. For this reason, I limited my library sequence diversity to 9216 sequences. This is 521 fold

smaller than Leulib, 228 fold smaller than Dawson et al, and 2.7×10^6 fold smaller than Herrmann et al, although just 15-fold smaller than the number of sequences that Herrmann et al report were able to insert into membranes.

4.4.2 Designing the interface and sequence context

I used a PCR scheme to generate a library with seven modestly randomized positions giving 9216 sequences that would include the combinations of interacting residues of the glycoporphin A and human/*Caenorhabditis elegans* BNIP3 type TMD dimer interfaces, see Figure 4.4. A polyleucine host background sequence was chosen in part to make this work directly comparable to previous library studies, while also maintaining high hydrophobicity necessary for membrane insertion and retention. The TMD length was kept to 16 amino acids, as shorter TMDs were shown to increase TOXCAT sensitivity in previous studies (Duong et al., 2007; Russ and Engelman, 1999)(Duong and MacKenzie, personal communication). This library is designed to test the relative importance of GxxxG motifs, GxxxG-like motifs (*i.e.* GxxxA, etc), and the presence of large and small polar residues to TMD dimerization.

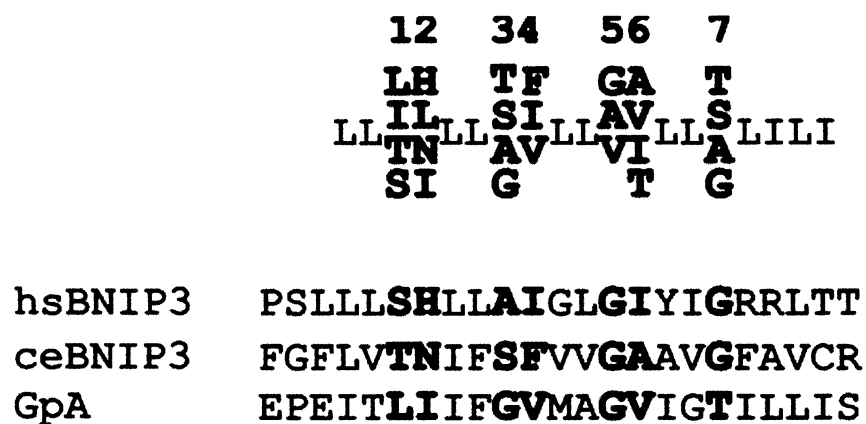


Figure 4.4 – PGM Library Design – The PGM small library design ($\sim 10^4$ sequences) is a combination of residues found at interacting positions of human BNIP3, worm BNIP3, and human GpA TMDs (bold) as well as other residues (top, orange) that were included to allow extra sequence diversity. Doing so allows me to ask which residue properties enhance or detract from TMD association: hydrophobic or hydrogen bonding (Pos 1 and 2), small apolar or hydrogen bonding (Pos 3 and 7), aromatic or beta branched (Pos 4), apolar of various sizes (Pos 5), and apolar of various sizes plus hydrogen bonding (Pos 6). PGM allows for GxxxG in single and tandem ‘glycine zipper’ varieties.

4.4.3 Degeneracy of the genetic code and amino acid choice

Several amino acid choices were incorporated that are not encoded by GpA or either human or worm BNIP3 parental sequences. In some instances, the degeneracy of the genetic code forced us to include more amino acids, and other times we intentionally included specific residues. Certain additional residues are included to ensure that we will be able to make a statistical determination on the probability of amino acid occurrence. If only two choices were allowed, and we find them in equal proportions, we are drastically limited on our descriptions of possible correlations. Either both amino acids are contributing equally to dimerization, or both are occurring at random. Often this simple example would be two similar amino acids (polar, small, etc.) due to the genetic code; an extreme example would be position 5, which is glycine in all three parental sequences. If we generate the library with only glycine at this position, we cannot determine the relative importance of glycine compared to any other residue. We include alanine and valine at this position in the hope that we will be able to determine experimentally the selection strength at which glycine and alanine each support dimerization, and whether glycine, alanine, or valine can be present in sequences that associate modestly or very tightly. By including more choices, we are allowing differentiation of effects by amino acid classes and at the same time allow statistical analysis to determine over- or under-represented effects. Therefore at positions where no residue is a clear ‘winner’ after selection, I am satisfied that a significant range of amino acid properties is being sampled with little or no effect on the degree of dimerization exhibited by the selected sequences.

4.4.4 Fusion protein expression level and antibiotic selection considerations

For very strongly associating TMDs, the TOXCAT interaction assay is insensitive to moderately disruptive mutations (Lawrie et al., 2010). This effect is related to the expression level of the fusion protein: for some TMDs, this level of protein effectively saturates the membrane environment with fusion constructs, pushing the thermodynamic equilibrium to the dimer state. I alleviate this problem by using a modified TOXCAT plasmid containing a point mutation in the putative ribosome binding site (Jaszewski and MacKenzie, unpublished). The modified construct, RBS1, results in decreased expression of the ToxR-TMD-MBP fusion protein and thus causes a lower expression of reporter gene CAT activity. The design and analysis of two libraries PGM (traditional TOXCAT plasmid) and PGM-Low (RBS1 plasmid) has been used to develop the first description of residue combinatorial effects on TMD dimerization.

In previous TOXCAT studies, the typical cutoff for selection strength was 400 $\mu\text{g/ml}$ chloramphenicol (CAM) (Russ and Engelman, 2000). We found we could generate sequences in the traditional TOXCAT construct that survived at up to 500 $\mu\text{g/ml}$ and potentially even higher, but working at high CAM concentrations is difficult and so the RBS1 plasmid was employed to shift the stringency scale. By expressing less of the fusion protein, clones that would have survived at 400 $\mu\text{g/ml}$ now survive at 200 $\mu\text{g/ml}$; fewer than 0.2% of RBS1 clones survive at 300 $\mu\text{g/ml}$ CAM. I built the same library into both the standard and the RBS1 TOXCAT vector and selected sequences from each to correspond to the top 30%, 10%, 3%, and 1% of associating clones. Doing so allowed me to compare the libraries to one another and to ensure that a top tier strongly associating class of sequences was not being overlooked.

4.4.5 Statistical analysis

In this thesis, I use the hypergeometric function to calculate the significance of finding particular residues (or pairs of residues) in populations of sequences from unbiased and selected data sets. The hypergeometric function is a mathematical tool that computes the likelihood of obtaining at random a particular set of events from a known parental distribution of event probabilities. For instance, this approach is useful for calculating the chance of obtaining exactly three clubs in drawing five cards from a deck of 52 playing cards. The hypergeometric formula obtains this probability based on the parent population size (52 total cards in a deck), the number of successes in the parent population (13 cards are clubs), the sample size under consideration (5 cards to be drawn), and the number of successes to be obtained in the sample (draw exactly 3 clubs) using the formula for the Hypergeometric Equation:

$$P(X = k) = \frac{\binom{m}{k} \binom{N-m}{n-k}}{\binom{N}{n}} \quad \text{where} \quad \begin{array}{l} k \text{ is the \# of successes in the sample; 3 clubs} \\ n \text{ is the size of the sample; draw five cards} \\ m \text{ is the successes in the parent population; 13 clubs} \\ N \text{ is the parent population size; 52 cards} \end{array}$$

In this example, the calculated probability ($P=0.0815$) indicates that an individual drawing five cards from a deck will receive three clubs 8.15% of the time – such a hand could easily happen by chance. Drawing five clubs in a hand of five is more rare ($P=0.000495$); the calculated P value can serve as a guide to identify events that are likely to be non-random, such as your poker opponent randomly dealing five clubs to himself on three successive hands. In my results, I exploit the computing power of Microsoft Excel and its HYPERGEOM function to carry out my hypergeometric calculations.

For my library work, the ‘events’ are sets of TMD amino acid sequences (approximately 9216 unique clones): I have sequenced extensively from the unselected libraries (about 66 clones), and from colonies isolated at different antibiotic concentrations (about 66 clones). The hypergeometric function was used to determine if the unselected library contains bias relative to the theoretical design of the library. The function counts the occurrences of each possible residue and compares them to the expected random rate. I am then able to determine if results obtained in my experiments contain statistically over- or under-represented amino acid occurrences.

By sequencing many unselected clones from the library, I also use the hypergeometric function to account for differences in the known pooled data (unselected) compared to a hypothetical library set where every sequence is made in equal proportions. Knowing the bias in the libraries allowed me to predict whether all of the possible sequences that were designed have likely been captured in the library and to determine the significance of residue distributions seen under selection conditions relative to the level of bias that is actually present in the library.

4.4.6 Summary/conclusions

My thesis work revolves around the analysis of a small TMD association library (<10,000 sequences). Contained in this library are motifs from the strongly homodimeric GpA, human BNIP3, and *C. elegans* BNIP3 TMDs. The small library design implemented here allows for the analysis of weak to strong dimerization levels. The resulting data gives a detailed look at the ‘sequence walk’ required to produce strongly dimeric TMDs by comparing the relative significance of glycine, large/small polar, and

hydrophobic residues. The importance of identities at single positions in this library is presented in Chapter 5. Sufficient sequences were collected to allow for analysis of positional pairwise correlations, and Chapter 6 presents analysis and conclusions about combinatorial effects between sequence elements.

Chapter 5 Plating Experiments and Single Site Analysis

5.1 Plating PGM and PGM-Low reveals a range of dimerization propensities

5.1.1 Surviving colonies drop off exponentially with increasing chloramphenicol

To determine the dimerization tendencies of sequences encoded by the PGM and PGM-Low libraries (see section 4.3.4), I transformed *E. coli* NT326 cells (Russ and Engelman, 1999, 2000) with each library as described in Chapter 8 and plated serial dilutions on increasing concentrations of CAM to determine the number of surviving cells. Small aliquots of each transformation were plated within 20 minutes of transformation on CAM-free plates to estimate the number of unique transformants and ensure that the library was oversampled; for the purposes of these plating experiments, 30,000 unique transformants was considered adequate sampling of the libraries. Each TOXCAT construct encodes a ToxR-TMD-MBP fusion protein that is expressed and inserted across the inner membrane, and when homodimerization of the TMD brings the ToxR DNA binding domains into close proximity, they activate a promoter that drives the production of the reporter gene CAT. By selecting a library against increasing concentrations of CAM, I selected for cells at tiers of CAT concentration. The greater antibiotic resistance is inferred to arise from increased dimerization propensity of the ToxR-TMD-MBP fusion protein. (Note that for the same TOXCAT clone, small stochastic variations in the expression level of ToxR-TMD-MBP fusion protein can also have an effect on the total amount of CAT produced (Duong et al., 2007), which somewhat weakens the correlation of dimerization strength and CAT production.)

As shown in Figure 5.1, fewer colonies are viable as CAM concentration increases, and this manifests as an exponential decrease in surviving cells. The Log-linear drop off observed is very reproducible and is reminiscent of the exponential drop off reported by Russ et al (Russ and Engelman, 1999, 2000), although with a less steep slope. Because the ability to survive at higher levels of CAM is primarily the effect of an increased stringency for TMDs to form dimers, these plating experiments show that by isolating colonies selected at a given level of CAM I can obtain the clones with the tightest associating TMDs. Thus, by tuning the CAM concentration, I can select and sequence

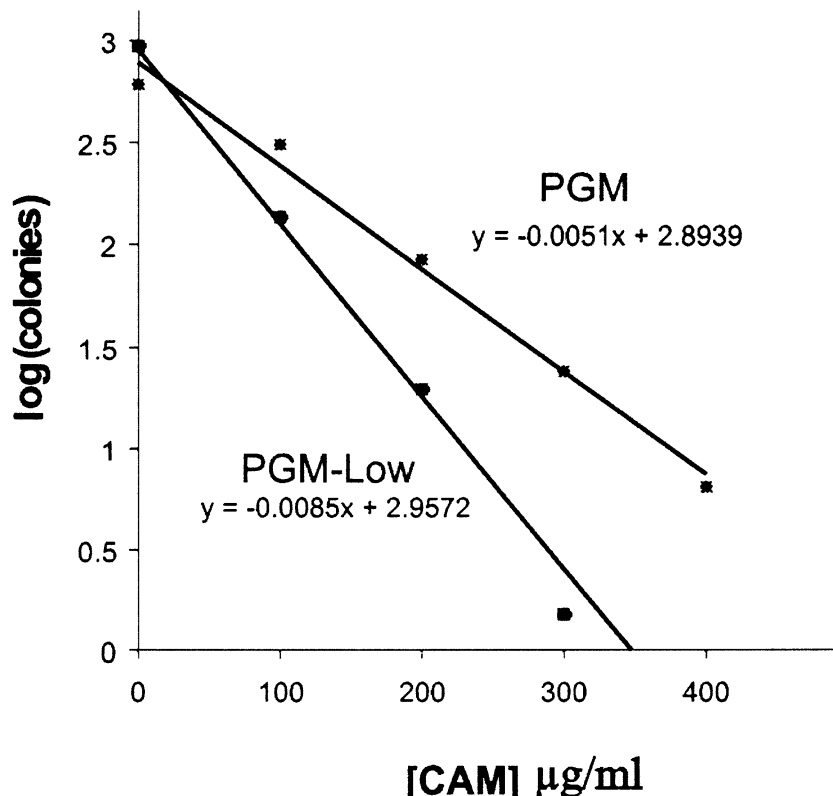


Figure 5.1 – Log-Linear decline – Surviving PGM and PGM-Low colonies drop off exponentially with increased chloramphenicol, as revealed when the log of surviving colonies is plotted against the drug concentration in the plates.

the top ~30%, ~1%, or any other arbitrarily small portion of the TMD sequences encoded by each library (see Table 5.1).

Table 5.1 - Surviving Fraction
[CAM] ($\mu\text{g/ml}$)

surviving fraction	PGM	PGM- Low
1%	400	200
3%	300	150
10%	200	100
30%	100	50

This procedure establishes the rank order of the average dimerization strength of clones on plates of different CAM concentrations, but not every clone from a high CAM plate will associate more tightly than every clone from a lower CAM plate because the clones that constitute the “1%” class are a subset of the “30%” class – exhaustive sequencing of the 30% pool should identify all of the clones present in the 1% pool.

5.1.2 ToxR-TMD-MBP fusions are properly inserted in bacterial inner membranes

To ascertain the cellular localization of the ToxR-TMD-MBP fusion proteins, I chose unique PGM and PGM Low clones at random from all CAM levels surveyed and assayed for maltose complementation. Only the sequences that are expressed and incorporated properly into the *E. coli* inner membrane will present MBP in the periplasm and confer the ability to grow on maltose plates to NT326 cells, which lack endogenous MBP (*malE*). Each clone was picked from a glycerol stock and grown overnight in M9 minimal media with glucose as the sole carbon source. Five micro-liters of the overnight culture was dotted on maltose plates (Russ and Engelman, 1999), which are made with M9 minimal media and maltose as the sole carbon source.

From the total of 330 total clones isolated from the PGM library, 99 unique sequences were tested on maltose-only media. Of these, a single sample (PGM 300- 30) failed to grow on a maltose minimal media plate. This sequence was removed from subsequent analysis, but the low incidence of poorly behaved sequences indicates that the vast majority of PGM clones are properly inserted into the inner membrane. The PGM-Low collection contains 337 total sequences, and I tested ninety-seven unique PGM-Low clones for maltose complementation. Two clones that were not under selection (0 µg/ml CAM) were unable to grow when maltose was the only available carbon source, but all clones from selected plates complemented *malE*. From these data, I conclude that the vast majority of both PGM and PGM-Low sequences are transcribed, translated, and inserted across the membrane correctly, and more importantly that the small fraction of constructs that behave poorly in this assay are not enriched by selection against CAM.

Accordingly, my data are not skewed by the presence of mis-localized fusion proteins.

The total number of clones sequenced at each concentration is listed in Table 5.2.

Table 5.2 - Summary of Library Sequences.

PGM				PGM-Low			
CAM μg/ml	Seqs	Acquired	Unique Seq	CAM μg/ml	Seqs	Acquired	Unique Seq
0	58		46	0	63		54
100	70		49	50	71		41
200	72		51	100	72		29
300	71		50	150	64		36
400	59		40	200	67		35
	330	Total	236		337	Total	195
	66	Avg.	47.2		67.4	Avg.	39
			Avg.				Avg.

5.1.3 PGM and PGM-Low constructs display fast and slow growth on maltose media suggesting a qualitative difference in dimer stability

Careful examination of the *malE* complementation plates revealed an unexpected secondary phenotype. I found that it could take up to six days for dotted cultures to grow to mature, visible colonies. However, some specimens matured in two or three days. When variable rates of growth on maltose plates have been seen by others in our research group, rapid growth has tended to correlate with weak ToxR-TMD-MBP dimerization. We hypothesize that this variable ability to uptake maltose is related to the availability of the maltose binding protein (MBP) domain of the fusion protein: perhaps strong TMD association makes MBP less sterically accessible, slowing the rate of maltose uptake. Slow growing clones likely have most of their fusion proteins in the dimeric state, which somehow impedes the ability of MBP to bind and transport maltose.

5.2 Sequencing reveals a mixture of sequence biases at different selection strengths

I sequenced clones from plates at different chloramphenicol (CAM) concentrations in order to determine how TMD sequence features changes with increasing selection strength, and I analyzed the same library sequences in two different vectors (TOXCAT and RBS1-TOXCAT) in order to determine the effect of altering the level of ToxR-TMD-MBP fusion protein expression. The CAM levels were chosen so that the most stringent plates would allow about 1% of the total library to survive (refer to Figure 5.1 and Table 5.1). From the plating experiments described above, I expect that the chosen selection strengths for PGM-Low (50-200 µg/ml CAM , in steps of 50 µg/ml)

will give surviving fractions of the library very similar to those obtained for the PGM selection strengths (100-400 $\mu\text{g/ml}$ CAM, in steps of 100 $\mu\text{g/ml}$). If this rank ordering of the clones by strength of TMD dimerization is consistent across the two vectors, then the sequences found in (for instance) the 3% pools from the two vectors should be quite similar.

From the PGM Library, I sequenced extensively from the 0, 100, 200, 300, and 400 $\mu\text{g/ml}$ CAM plates, and from the PGM-Low Library I sequenced from 0, 50, 100, 150, and 200 $\mu\text{g/ml}$ CAM plates. I obtained clean sequence for an average of sixty six sequences per CAM level in both libraries, which was more than sufficient to obtain meaningful P-values for residue biases at single positions using the hypergeometric function.

The relative frequency of amino acids by position is presented graphically in Figure 5.2. The most apparent of these is the position 5 glycine, which is the only residue to completely dominate the population of any one position in PGM or PGM-Low. Phenylalanine, however, makes up three-quarters of the population of position 4 at selection levels PGM-400 and PGM-Low 200. These trends are easy to see from a visual representation, but other trends are more difficult to glean because of bias. The parental library, represented by the residue frequencies at PGM 0 and PGM-Low 0, are not the 'Ideal' distribution I designed my experiments to create (refer to Figure 5.2). Strong bias exists at position 7 against serine and at position 6 against isoleucine, which were absent from the unselected PGM clones. Bias in the original PCR product causes these residues to occur at such low frequencies that they are missed at the unselected level, although

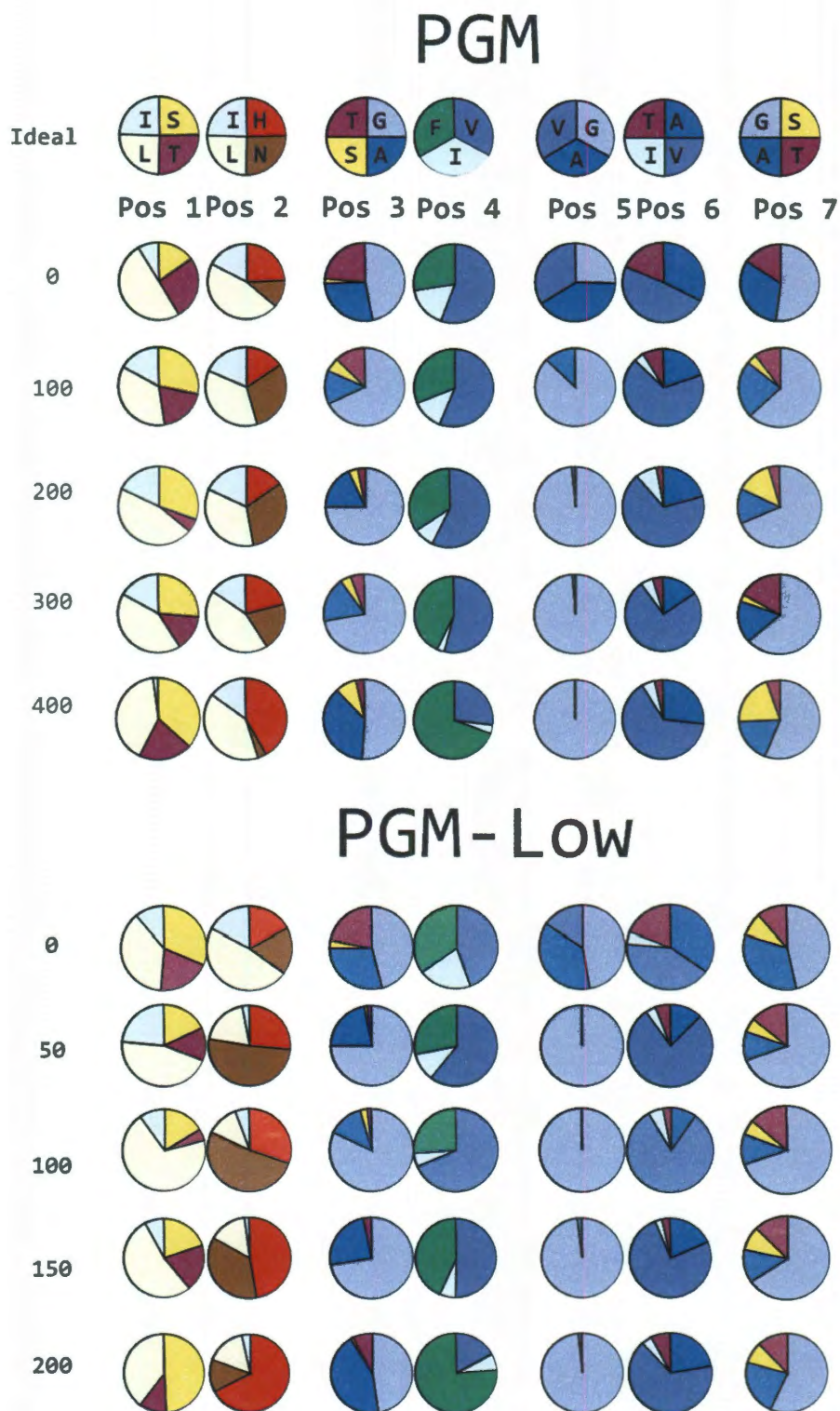


Figure 5.2 - PGM/PGM-Low single site residue frequencies – Fraction of residues seen at each selection strength (left) and at each position are shown graphically. PGM and PGM-Low unselected sequences are biased from the 'Ideal' distribution (top) and strong trends for positions 4 and 5 are seen as [CAM] increases.

their appearance in PGM Low unselected sequences shows that they are indeed present (the two libraries were clones from the same PCR product). The selective pressure is so strong that serine appears at higher CAM levels as a yellow slice (see Figure 5.2), demonstrating the power of the selection approach to enrich the surviving clones with particular residues. The PCR cloning method is likely the source of unselected bias and a statistical analysis of this bias is presented in Table 5.4. This bias is taken into account when computing single site amino acid P-values observed at different selective pressures. Rather than calculating probabilities as if sequences were drawn from the intended library, I use the experimentally observed sequence bias present in the unselected library as the ‘null hypothesis’. These P-values are presented in Tables 5.5 and 5.6 simultaneously with odds/ratios. These values are discussed in detail in the following section 5.2.1.

Of the three sequences that were initially used to design PGM and PGM-Low only GpA is found a single time in winning sequences, defined as those identified from each library at the highest drug concentration, see Table 5.3. Several sequences are found in both PGM and PGM-Low winners. I interpret this to mean that selection in both libraries is reporting on the same phenomenon: drug resistance is resulting from sequence-specific TMD-mediated dimerization. The sequences of all clones obtained from plates at all CAM concentrations and all P-values are presented in the Appendices, Tables 1.1 and 1.2, respectively.

Table 5.3 - Winning Sequences - The top 1% PGM and PGM-Low sequences.

PGM 400		PGM-Low 200	
LLSILLGVLLGILLS	LLSHLLAVLLGALLA	LLSLLLAFLLGVLLA	LLLNLLGFLLGVLLG
LLTHLLSVLLGVLLG	LLSHLLAFLLGALLA	LLSHLLAFLLGVLLS	LLSHLLGFLLGVLLG
LLTHLLSVLLGVLLG	LLTLLLGFLLGALLT	LLHLLGVLLGVLLG	LLHLLGFLLGVLLG
LLLLLLAFLLGVLLG	LLSHLLAVLLGILLA	LLHLLGVLLGVLLG	LLSHLLAFLLGVLLS
LLHLLGILLGVLLG	LLTHLLGVLLGVLLG	LLSILLGFLLGVLLG	LLHLLGFLLGVLLG
LLTHLLAILLGALLA	LLSHLLTFLLGALLS	LLSLLLAFLLGALLA	LLSHLLAFLLGVLLA
LLSHLLGFLLGALLS	LLIILLGFLLGVLLS	LLLNLLGVLLGVLLG	LLSHLLAFLLGALLA
LLHLLGFLLGALLS	LLTLLLAFLLGVLLG	LLLNLLGILLGILLG	LLSHLLAFLLGALLA
LLLNLLGVLLGVLLG	LLSLLLAFLLGALLA	LLSLLLAFLLGVLLA	LLSHLLAFLLGTLG
LLLLLLGVLLGVLLG	LLSHLLAVLLGALLT	LLTHLLAFLLGTLA	LLSHLLAFLLGALLT
LLSHLLAVLLGALLG	LLSLLLSFLLGALLA	LLSHLLAFLLGALLA	LLTLLLGVLLGVLLS
LLSLLLGFLLGVLLG	LLTHLLSFLLGALLS	LLLNLLGVLLGVLLG	LLSHLLAFLLGALLA
LLSHLLAFLLGALLA	LLSLLLAFLLGALLA	LLSHLLSVLLGALLA	LLHLLGVLLGVLLG
LLSLLLAFLLGALLA	LLSLLLAFLLGALLS	LLLLLLGFLLGVLLT	LLLNLLGFLLGVLLG
LLIILLGFLLGVLLG	LLSILLGFLLGVLLG	LLTHLLTFLLGALLT	LLTHLLAVLLGVLLA
LLSLLLAFLLGVLLG	LLTHLLGVLLGVLLG	LLSHLLAFLLGALLA	LLTHLLAFLLGALLG
LLTHLLTFLLGTLG	LLLLLLAFLLGVLLG	LLLNLLGFLLGVLLG	LLHLLGVLLGVLLG
LLIILLGFLLGVLLS	LLHLLGFLLGVLLG	LLSHLLTFLLGILLG	LLHLLGILLGVLLG
LLLLLLAFLLGVLLG	LLIHLLSFLLGTLLG	LLSHLLAFLLGALLA	LLSHLLAFLLGALLG
LLLLLLAFLLGVLLG	LLSILLGVLLGVLLG	LLSLLLGVLLGVLLS	LLSHLLTFLLAVLLT
LLHLLGFLLGVLLG	LLSHLLGFLLGVLLG	LLSHLLAVLLGVLLT	LLSLLLAFLLGVLLG
LLIILLGFLLGVLLS	LLHLLGFLLGVLLG	LLHLLGFLLGVLLG	LLHLLGFLLGVLLG
LLSLLLAFLLGVLLA	LLIILLAFLLGVLLG	LLHLLGFLLGVLLG	LLHLLGVLLGVLLG
LLTLLLAFLLGVLLG	LLSLLLAFLLGVLLG	LLTHLLAFLLGALLG	LLSLLLAFLLGALLT
LLTLLLGFLLGVLLG	LLLLLLGFLLGVLLS	LLHLLGFLLGVLLG	LLLNLLGILLGVLLG
LLLLLLAFLLGVLLG	LLSNLLGFLLGVLLG	LLSHLLTFLLGVLLG	LLHLLGFLLGVLLG
LLHLLGFLLGVLLG	LLIILLGVLLGVLLT	LLSHLLAFLLGTLA	LLTHLLAFLLGTLG
LLLLLLGFLLGVLLA	LLLLLLGFLLGVLLS	LLSLLLAFLLGVLLG	LLLNLLGILLGVLLG
LLHLLGVLLGILLG	LLTHLLGVLLGVLLG	LLSILLAFLLGVLLG	LLSHLLTFLLGILLG
		LLLNLLGFLLGVLLG	LLHLLGFLLGVLLG
		LLSLLLGFLLGVLLT	LLSLLLAFLLGALLT
		LLHLLGFLLGVLLG	LLSHLLAFLLGTLA
		LLHLLGFLLGVLLG	LLHLLGFLLGVLLG
		LLSHLLAFLLGVLLS	

Using the observed residue frequencies at each position, I can calculate the significance (as P-values) of the chance of pulling my unselected residue distributions from a hypothetical non-biased parental set (Table 5.2). I find that PGM and PGM-Low libraries contain several strongly biased positions ($<10^{-2}$), which appear to be due to biases in the original PCR product from either the oligonucleotide synthesis or PCR effects. The most severely biased positions in PGM are Position 6 (biased against isoleucine) and Position 7 (biased against serine); statistically significant biases also occur at these positions for the residues that are over-represented relative to the ideal

situation. Although the unselected library shows bias, I can properly account for this bias in subsequent analysis by using the observed unselected ratios as the null hypothesis.

Table 5.4 - P values for unselected residues arising from the intended library by chance.

PGM Bias

Pos 1	L	I	S	T
	2.4×10^{-5}	1.0×10^{-3}	3.0×10^{-2}	1.2×10^{-1}
Pos 2	L	I	H	N
	1.9×10^{-4}	5.0×10^{-2}	1.2×10^{-1}	7.7×10^{-3}
Pos 3	G	A	S	T
	1.9×10^{-4}	1.1×10^{-1}	1.0×10^{-6}	1.2×10^{-1}
Pos 4	F	I	V	
	1.1×10^{-1}	5.0×10^{-2}	6.3×10^{-7}	
Pos 5	G	A	V	
	1.2×10^{-1}	5.2×10^{-3}	3.0×10^{-2}	
Pos 6	A	V	I	T
	4.6×10^{-2}	6.9×10^{-5}	5.4×10^{-8}	7.3×10^{-2}
Pos 7	G	A	S	T
	7.5×10^{-6}	4.6×10^{-2}	5.4×10^{-8}	3.0×10^{-2}

PGM-Low Bias

Pos 1	L	I	S	T
	7.4×10^{-3}	3.3×10^{-3}	5.2×10^{-2}	6.7×10^{-2}
Pos 2	L	I	H	N
	5.4×10^{-5}	4.7×10^{-2}	4.7×10^{-2}	4.7×10^{-2}
Pos 3	G	A	S	T
	1.4×10^{-4}	9.0×10^{-2}	2.8×10^{-6}	1.1×10^{-1}
Pos 4	F	I	V	
	2.2×10^{-2}	8.8×10^{-2}	3.6×10^{-4}	
Pos 5	G	A	V	
	5.4×10^{-5}	1.3×10^{-2}	2.9×10^{-2}	
Pos 6	A	V	I	T
	2.2×10^{-2}	1.8×10^{-3}	1.9×10^{-5}	6.7×10^{-2}
Pos 7	G	A	S	T
	1.4×10^{-4}	2.5×10^{-2}	1.2×10^{-3}	2.3×10^{-3}

Table 5.5 - Single site residue biases in PGM under selection
Over represented

		CAM (µg/ml)			
		100	200	300	400
Position 5	G	6.0×10^{-27}	9.4×10^{-41}	3.6×10^{-40}	2.0×10^{-35}
Position 4	F			1.4×10^{-3}	1.9×10^{-11}
Position 3	G	2.4×10^{-4}	5.7×10^{-7}	8.9×10^{-6}	2.9×10^{-3}
	S				
Position 6	V	6.1×10^{-4}	3.1×10^{-4}	3.5×10^{-6}	4.8×10^{-3}
	I		6.9×10^{-3}		
Position 2	H				1.2×10^{-3}
	N	3.4×10^{-5}	4.9×10^{-6}		
Position 1	I		4.9×10^{-3}		
	S	2.2×10^{-3}	5.9×10^{-4}		3.5×10^{-5}
Position 7	G		1.9×10^{-3}		
	S				2.3×10^{-5}

Under represented

		CAM (µg/ml)			
		100	200	300	400
Position 5	A	6.1×10^{-7}	6.5×10^{-15}	1.1×10^{-14}	1.0×10^{-13}
	V	1.2×10^{-13}	5.2×10^{-14}	8.0×10^{-14}	1.3×10^{-11}
Position 3	A	3.9×10^{-3}			
	T		5.6×10^{-7}	2.9×10^{-5}	1.4×10^{-5}
Position 4	I			1.5×10^{-4}	1.0×10^{-3}
	V				8.0×10^{-6}
Position 6	A	7.0×10^{-3}	9.2×10^{-3}	5.3×10^{-4}	
	T	8.6×10^{-3}	3.6×10^{-5}	2.3×10^{-4}	3.7×10^{-4}
Position 1	T		6.4×10^{-6}		
Position 7	A		1.5×10^{-4}	1.3×10^{-3}	
	T		1.9×10^{-3}		

Table 5.6 - Single site residue biases in PGM-Low under selection.

Over represented

		CAM (µg/ml)			
		50	100	150	200
Position 5	G	1.6×10^{-26}	6.0×10^{-24}	1.6×10^{-19}	1.8×10^{-20}
Position 4	F				4.9×10^{-12}
	V	2.3×10^{-3}	2.7×10^{-5}		
Position 3	G	6.0×10^{-7}	2.7×10^{-10}	1.6×10^{-5}	3.6×10^{-3}
	A				
Position 6	V	4.0×10^{-10}	1.2×10^{-12}	3.1×10^{-8}	3.0×10^{-5}
Position 2	H				2.9×10^{-19}
	N	1.2×10^{-10}	3.9×10^{-11}	2.0×10^{-4}	
Position 1	L		4.6×10^{-8}		
	I	1.0×10^{-3}			1.1×10^{-3}
	S				
Position 3	7	4.7×10^{-5}	3.1×10^{-5}	7.0×10^{-4}	

Under represented

		CAM (µg/ml)			
		50	100	150	200
Position 5	A	8.6×10^{-15}	5.4×10^{-15}	2.1×10^{-13}	2.1×10^{-12}
	V	4.5×10^{-6}	3.8×10^{-6}	1.8×10^{-4}	9.0×10^{-6}
Position 3	A		1.7×10^{-3}		
	T	3.4×10^{-7}	2.7×10^{-7}	1.6×10^{-5}	8.7×10^{-4}
Position 4	I		2.7×10^{-4}	1.1×10^{-3}	6.4×10^{-4}
	V				3.1×10^{-6}
Position 6	A	1.5×10^{-5}	6.5×10^{-7}	2.1×10^{-3}	2.1×10^{-3}
	T	8.9×10^{-4}	3.4×10^{-5}	1.5×10^{-4}	4.9×10^{-3}
Position 1	I				3.6×10^{-4}
	S	4.6×10^{-3}	1.8×10^{-3}		
	T		1.9×10^{-4}		
Position 7	A	1.2×10^{-5}	9.3×10^{-6}	9.0×10^{-5}	
Position 2	L				6.4×10^{-8}
	T				3.5×10^{-4}

5.2.1 PGM single site residue trends reveal a positional hierarchy underlying self association of BNIP3-like TMDs

Each degenerate position in my library design has amino acids that were significantly over- or under-represented at one or more selection strengths. I present these according to the rank of the associated P-value calculated relative to a non-selected population, *i.e.*, the likelihood that the observed amino acid frequency could occur at random from the un-selected parent population. Table 5.5 contains all the calculated P-values for the PGM library, and a binned, graphical representation of the most significant P-values at each position as a function of chloramphenicol concentration is presented in Figure 5.3, in which panel A contains the over-represented residues and B contains the under-represented residues.

The strongest observed bias for an amino acid in tightly associating TMD sequences is for glycine at position 5, which is a glycine in all three parental GpA, human BNIP3, and worm BNIP3 TMDs. Glycine is significantly over-represented at this position in all PGM sequences obtained under selection. Although strongly selected against at high CAM concentrations, alanine can be found at this position in sequences derived from 100 $\mu\text{g/ml}$ CAM plates. Valine, in contrast, is strongly selected against at position 5 at all selection strengths. We infer that a position 5 glycine is necessary for forming the strongest TMD dimers possible within the sequence constraints of the PGM library. From Figure 5.1, we can see that the clones that survive at 100 $\mu\text{g/ml}$ correspond to approximately the 30% most tightly associating library sequences. The essentially complete elimination of valine and alanine from this set of sequences shows that the strongly polar histidine and asparagine residues at position 2 cannot drive even modest

TMD dimerization without a glycine at position 5. Glycine at this position is therefore part of a ‘motif’ needed for tight dimerization. I speculate that this likely is necessary to permit close approach of the two helix backbones, and that intermonomer non-canonical $C\alpha-H\cdots O=C$ hydrogen bonds probably form in the tightest interacting sequences.

Position 4 reveals a strong but not absolute bias for phenylalanine to be present in strongly associating sequences (400 and 300 $\mu\text{g/ml}$ CAM). Valine and isoleucine are modestly biased against at these higher CAM concentrations, but the indications of bias are not statistically significant for any residue at 100 or 200 $\mu\text{g/ml}$ CAM. All three residues at this position were present in the parental sequences; no additional choices were invoked in the library design. The data indicate that for sequences that survive at 200 $\mu\text{g/ml}$ CAM, which is the 10% most strongly dimerizing library sequences, the bias towards phenylalanine can hardly be detected. Once we look at the top 3% or 1% of sequences, a strong bias for phenylalanine is revealed, but sequences without a phenylalanine can still be found. I infer that the different parental amino acids optimize packing interactions, perhaps in a way that favors intermonomer non-canonical $C\alpha-H\cdots O=C$ hydrogen bond formation. Despite the strong statistical bias, it is clear that the phenylalanine at position 4 is not absolutely required for very tight dimerization. This position seems to provide a ‘sequence context’ that influences dimerization, rather than corresponding to a ‘motif’ requirement. Of course, the limited nature of our library did not allow us to fully sample the possible residues at this position, and some of the untested residues might be so strongly biased against that the others would constitute a broad consensus motif position.

Position 3 shows a complex set of moderate biases as the selection stringency changes. Glycine is over-represented in sequences under selective pressure except in the top winner pool, where it is unbiased. Alanine is biased against at low selection strengths, especially at 100 $\mu\text{g/ml}$ CAM, but is slightly favored at the highest selection level. Threonine is biased against at all selection levels, and strongly biased against at 200 $\mu\text{g/ml}$ CAM and above. Serine is slightly over-represented at low stringency and moderately over-represented at the highest stringency. No clear trends based on residue size or availability of a side chain hydroxyl can be established. It is likely that glycine and serine are able to support strong and specific interactions, but we cannot infer what the physical basis for these might be.

Position 2 exhibits a dramatic reversal of residue biases across different stringencies, see Figure 5.3A (top right). Asparagine is strongly over-represented at 100 and 200 $\mu\text{g/ml}$ CAM, is essentially unbiased at 300 $\mu\text{g/ml}$ CAM, and then is modestly biased against at 400 $\mu\text{g/ml}$ CAM. Histidine follows the opposite trend, being slightly biased against at 100 and 200 $\mu\text{g/ml}$ CAM, unbiased at 300 $\mu\text{g/ml}$ CAM, and modestly biased for at 400 $\mu\text{g/ml}$ CAM. The complexity of these results is rather unexpected given the straightforward rationale of our library design. We allowed two large hydrophobic residues (isoleucine and leucine) and two polar residues (histidine and asparagine) to appear in order to test the role of large polar residues in driving TMD dimerization. Leucines are slightly biased against at low CAM concentrations, but overall there is no significant bias for or against the aliphatic residues, indicating that either polar or aliphatic side chains are consistent with modest or strong dimerization. However, the excess of asparagine at low selective pressures and of histidine at high selective pressure

shows that these two residues influence dimerization quite differently. At low selection strength, perhaps the asparagine side chains are making symmetric Asn:Asn hydrogen bonds, which histidine side chains cannot do. Such hydrogen bonds could drive moderate levels of dimerization. At high selection strengths, side chain hydrogen bonds would need to be combined with excellent packing, and it is possible that the rest of the variable residues (or sequence context) are more compatible with the geometry of histidine-mediated tight dimerization than with asparagine-mediated tight dimerization.

At position 1, isoleucine and serine are slightly over-represented at the expense of leucine and threonine in sequences selected at 100 $\mu\text{g/ml}$ CAM. Whereas leucine remains slightly under-represented at all CAM levels, isoleucine is over-represented at up to 300 $\mu\text{g/ml}$ CAM but strongly under-represented at the highest selection strength. Threonine is slightly under-represented at high CAM levels and strongly under-represented at 200 $\mu\text{g/ml}$ CAM, and serine is modestly over-represented at intermediate CAM levels but strongly over-represented at 400 $\mu\text{g/ml}$ CAM. It is clear that polar residues are not required at this position for tight dimerization, but the depletion of isoleucine at the highest stringency level suggests that its bulky β -branched side chain is more difficult to accommodate in the interface of a strongly associating dimer than a leucine side chain. The bias for serine at high stringency could reflect BNIP3 type hydrogen bonding, in which a large polar side chain donates a hydrogen bond to a small polar side chain, but it would also be consistent with a role for a small residue or for a polar zipper.

Position 6 shows modest bias favoring valine and isoleucine at all selection levels, and alanine and threonine are modestly under-represented at all selection levels.

Variations in the degree of bias are too slight at this position to interpret trends, especially since the prevalence of isoleucine in the unselected sequences is very low, so I conclude that position 6 exhibits a modest preference for large hydrophobic residues over small or hydrogen bonding residues.

Like position 6, alanine and threonine are highly biased against at all CAM concentrations at position 7. Serine is over-represented at position 7 at all selection stringencies, but the significance of this is difficult to assess quantitatively: serine was not identified at this position in 58 unselected sequences, so the extent to which serine has been enhanced in the selected pools is not clear. Glycine is only slightly favored at position 7, and the statistical significance of this bias disappears at the highest selection strength. The depletion of threonine is at odds with the GpA association motif, **LlxxGVxxGVxxT**, which features threonine at this position.

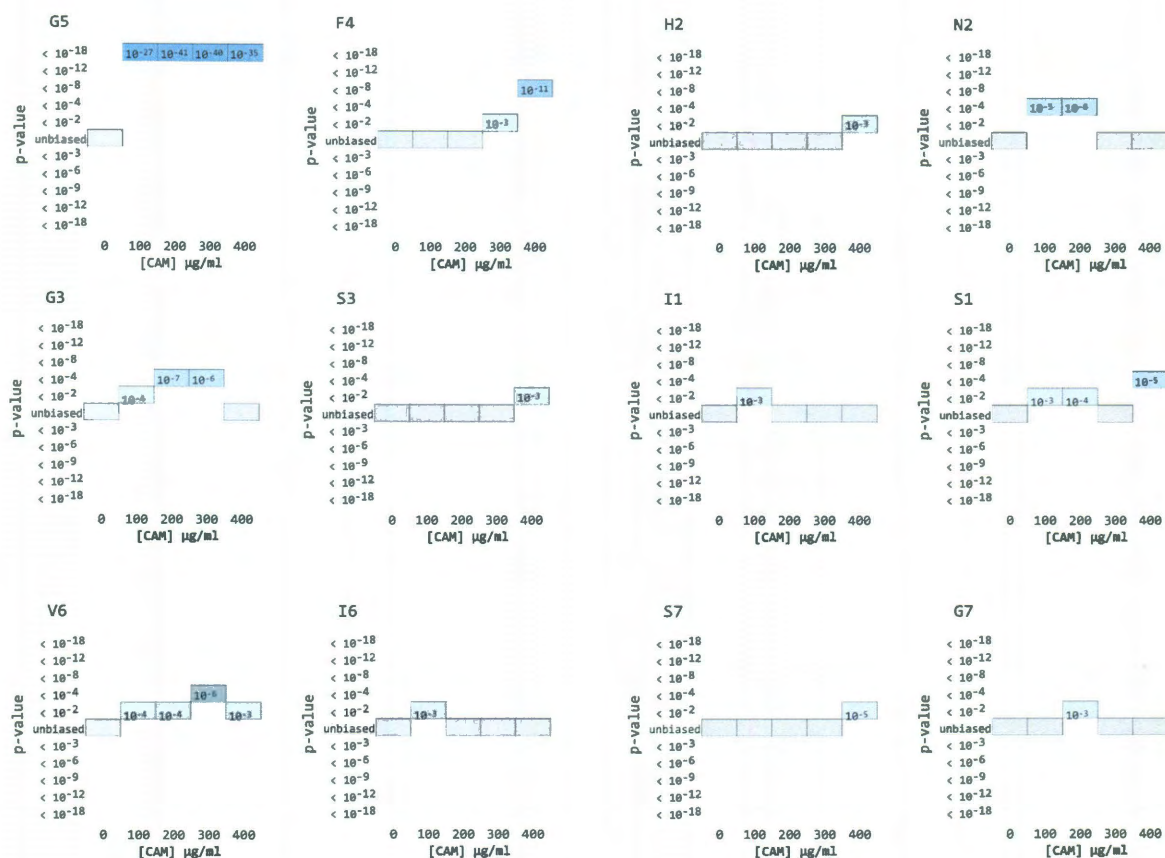


Figure 5.3A - PGM Over-represented single site residue biases – A binned, graphical representation of the P-values at each position as a function of chloramphenicol concentration is displayed. Blue bars above the unbiased level are depictions of over-represented residues; the higher (and darker) the bar, the greater the over-representation of that residue. P-values for these over-representations are given inside each box.

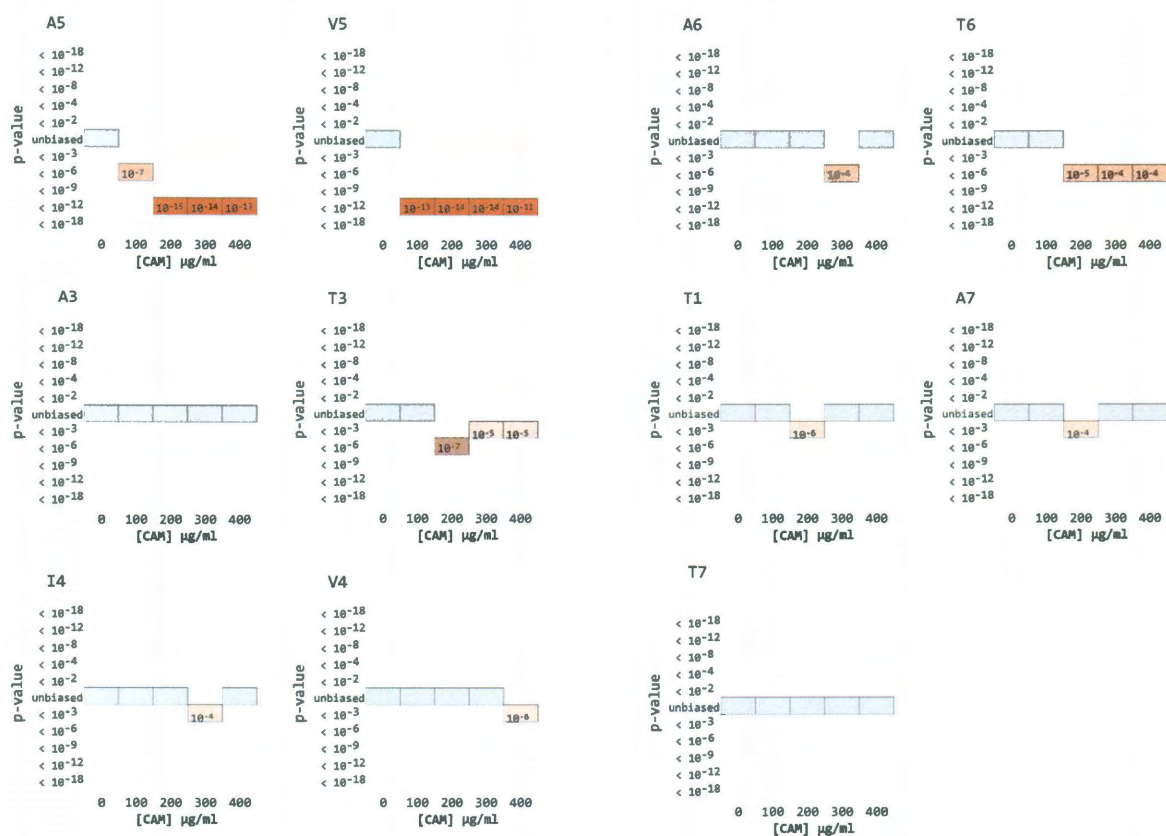


Figure 5.3B - PGM Under-represented single site residue biases – A binned, graphical representation of the P-values at each position as a function of chloramphenicol concentration is displayed. Orange bars below the unbiased level are depictions of under-represented residues; the higher (and darker) the bar, the greater the over-representation of that residue.

5.2.2 PGM-Low single residue trends

Here I describe PGM-Low single site residue biases reported in detail in Table 5.6 and portrayed graphically in Figure 5.4 while making appropriate comparisons to PGM.

Glycine at position 5 is the most favored residue in the most strongly associating set of PGM-Low sequences and at lower selection levels. This is the same effect we observed in PGM, and the strong similarities support the idea that the two libraries have very similar relative ranking of sequences.

Phenylalanine at position 4 dominates the top 1% of sequences in PGM-Low (200 $\mu\text{g/ml}$ CAM), as is the case in PGM. The degree of bias is about the same in both libraries, although at the second to top selection level (3%) a stronger bias for phenylalanine is seen in PGM-Low than in PGM. There is a very strong bias against valine at 200 $\mu\text{g/ml}$ CAM, but at the lower levels (100 and 150 $\mu\text{g/ml}$ CAM) valine is over-represented. No statistically significant bias of valine was found at low selection levels in PGM. Isoleucine is under-represented at this position at selection levels higher than 50 $\mu\text{g/ml}$ CAM.

At position 3, glycine is over-represented at all levels of PGM-Low except at the top 1%, where it is essentially unbiased, as was seen in PGM. Also as in PGM, threonine is under-represented at all levels and alanine is under-represented in the top 10% sequences. In PGM-Low, alanine is over-represented in the top 1%, whereas in PGM serine was over-represented. The close correspondence between PGM and PGM-Low for the most significant biases indicates that the two libraries rank sequences similarly, and suggests that differences seen here may be due to random chance.

At position 2, the histidine and asparagine effects seen with PGM-Low are nearly identical to those in PGM: asparagine is over-represented at low to moderate selection levels and histidine is over-represented in the top 1%. However, leucine and isoleucine are both biased against in the top 1% of PGM-Low, instead of being unbiased as was the case in PGM. This bias against the hydrophobic residues in PGM-Low is statistically significant, and could indicate that hydrogen bonding side chains are more important to the ability to drive strong dimerization when the TMDs in question are dilute.

At position 1 isoleucine is biased against at all selection levels except 150 $\mu\text{g/ml}$ CAM, whereas serine is over-represented at the highest selection level. Leucine and isoleucine are over-represented at 100 and 50 $\mu\text{g/ml}$ CAM, respectively. Isoleucine is biased against at 200 $\mu\text{g/ml}$ CAM. Serine and threonine are under-represented at low selection levels. The only commonality with PGM is that serine is overrepresented at the top selection level, and the most significant deviation from PGM is that leucine is very strongly over-represented at 100 $\mu\text{g/ml}$ CAM and moderately over-represented at 200 $\mu\text{g/ml}$ CAM in PGM-Low, whereas leucine is slightly biased against at these selection levels in PGM.

At position 6, isoleucine is neither over- nor under-represented but valine is strongly over-represented, as in PGM. Alanine and threonine are strongly under-represented at this position, much as in the PGM results.

At position 7, glycine is over-represented at all but the highest selection levels, while alanine is under-represented at all but the highest levels, much as is seen in PGM. Serine and threonine are not significantly biased in PGM-Low, whereas threonine is modestly biased against and serine is favored at the highest stringency in PGM.

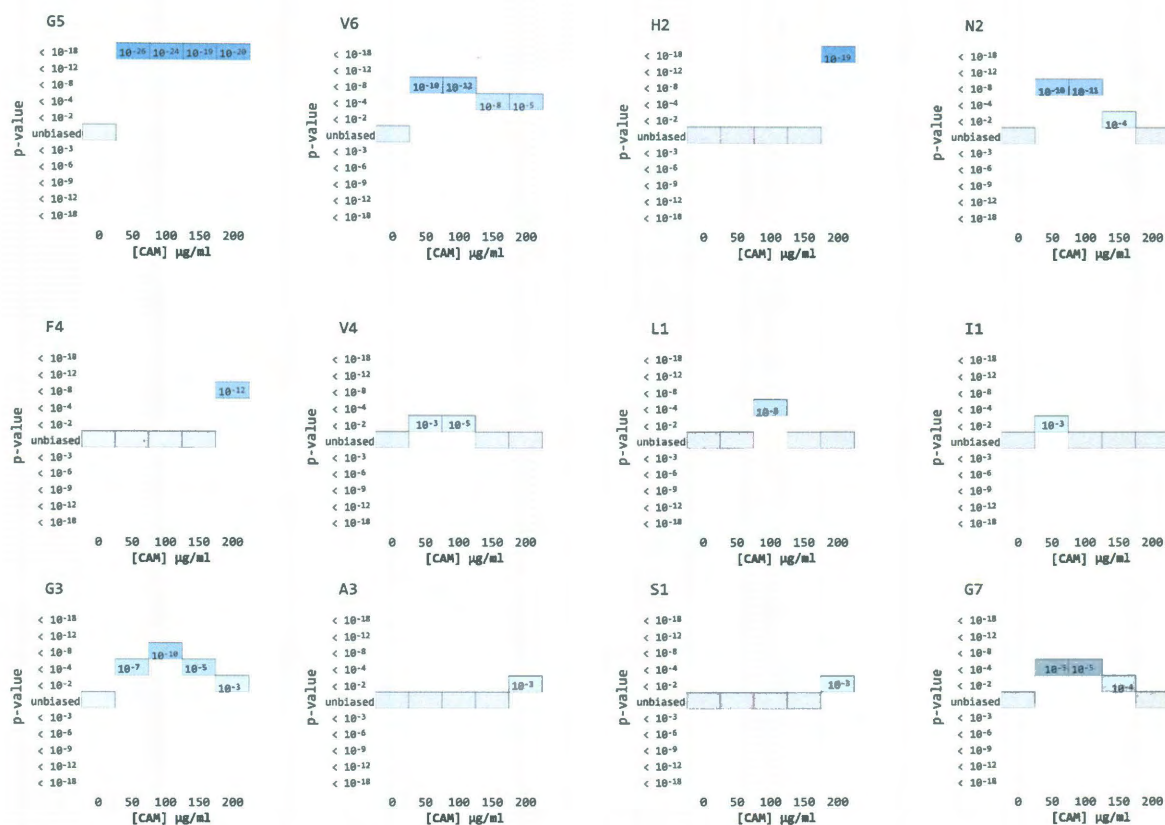


Figure 5.4A – PGM-Low Over-represented single site residue biases – A binned, graphical representation of the P-values at each position as a function of chloramphenicol concentration is displayed. Bars above the ‘unbiased’ level are depictions of over-represented residues, and the value within the box indicates the P-value associated with that degree of over-representation.

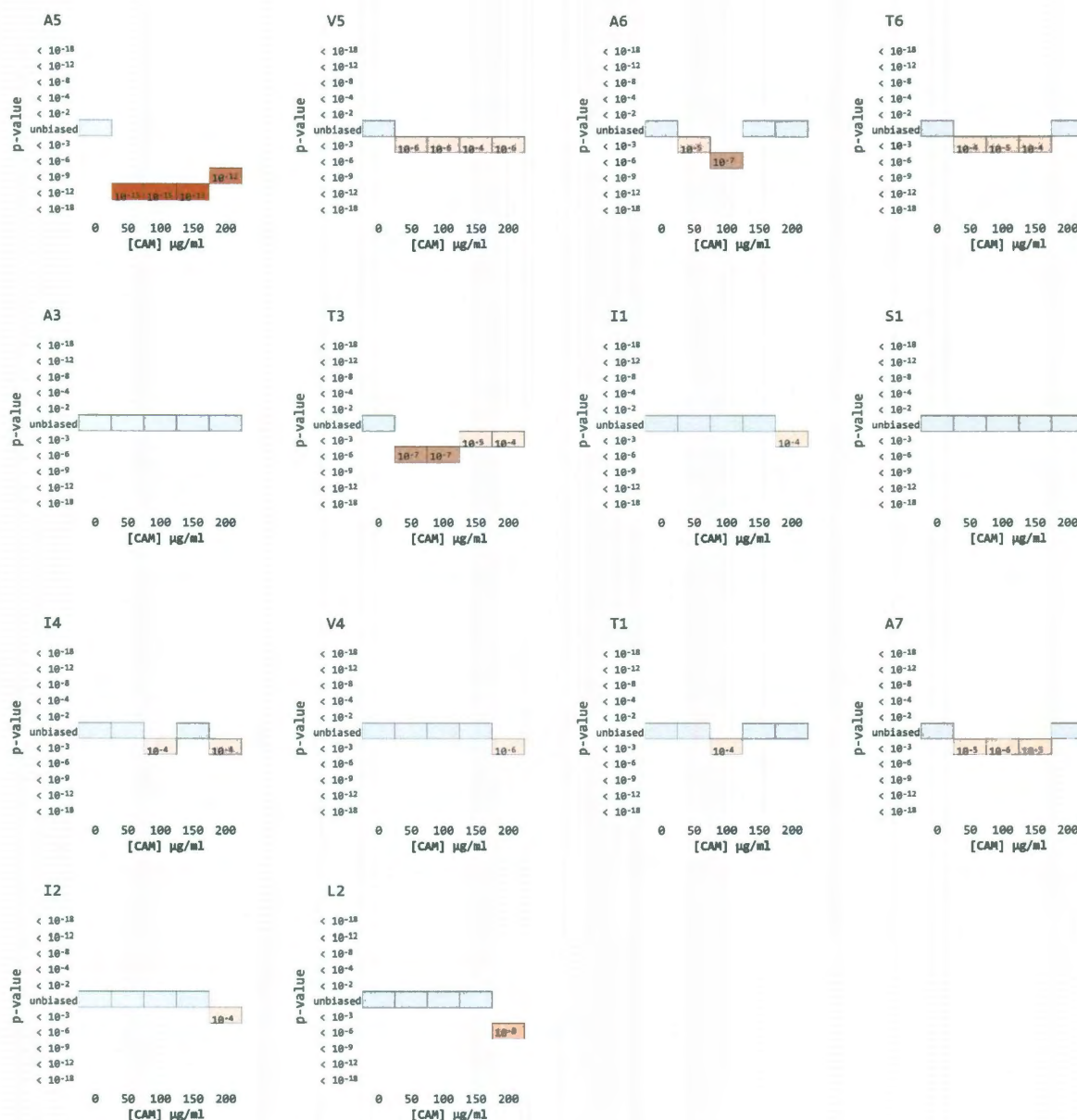


Figure 5.4B – PGM-Low Under-represented single site residue biases - A binned, graphical representation of the P-values at each position as a function of chloramphenicol concentration is displayed. Bars below the ‘unbiased’ level denote under-represented P-residues.

5.3 Comparison of PGM 400 and PGM-Low 200

5.3.1 Similarities/differences between winning sets

The winners of both libraries were defined as when 1% of total colonies plated were viable, which occurred at 400 and 200 $\mu\text{g/ml}$ CAM respectively for PGM and PGM-Low. All of these strongly dimeric TMDs contained a glycine at position 5 except a single PGM-Low isolate (SHxxTFxxAVxxT) where it is replaced by alanine. This glycine is common to all three parental sequences from which the library was designed. Although this central glycine is required to reach the strongest level of resistance to CAM, it is noteworthy that neither position 3 nor 7, the other positions that contain glycine in a parent sequence, show a bias towards (or against) glycines at the highest stringency level. A statistically significant bias towards glycine at position 5 can be seen for both libraries even at the lowest CAM selection level (refer to Table 1), indicating that glycine at this position is very important to even modest levels of TMD/TMD interaction. Position 2 shows a slight bias against the hydrophobic residues leucine and isoleucine but the proportion of polar TMDs that contain histidine or asparagine changes dramatically at different selection strengths. Histidine is essentially under-represented at less than the 1% stringency level, where it becomes significantly over-represented. Asparagine displays the opposite trend, being over-represented at lower selection levels then becoming under-represented at the 1% level. In describing possible motifs, histidine seems to be favored in strongly dimeric sequences, and asparagine is favored in moderately dimeric TMDs. Phenylalanine at position 4 and valine at position 6 are

strongly over-represented in PGM and PGM-Low winners and likely facilitate tight packing between strongly associating TMDs.

Although most biases in the two libraries are very similar (compare Figures 5.3 and 5.4, or Tables 5.5 and 5.6), there are exceptions where biases in PGM and PGM-Low pools in the winners do not agree. At position 3, serine is over-represented in PGM but not PGM-Low; at position 6, isoleucine is over-represented in PGM but not in PGM-Low; and at position 7, serine is over-represented in PGM but not PGM-Low. These differences may represent true differences between the behavior of TMD interactions in the two vectors, but given how closely the strongest biases are correlated between these two data sets, these differences may simply indicate the noisiness of library approaches to these types of questions. Conclusions that are common between the data sets and have very significant P-values can be interpreted as general findings, but the three differences between the data sets described have moderate P-values by hypergeometric analysis. Since any biases introduced at any stage in this analysis could skew the hypergeometric analysis somewhat, the lack of agreement for these more subtle biases between the data sets could be taken as an indication of the threshold at which our confidence in the statistical analysis should be questioned.

5.3.2 Selective pressure causes only slight differences between winning sets

The single site propensities of the top 1% PGM sequences were compared to those of the top 1% PGM-Low sequences using the hypergeometric function in order to determine the likelihood that the winners of PGM were being pulled from the same ‘pool’ that contained the PGM-Low winners. The single residue P-value analysis (Table 5.7)

shows that the majority of positions display no significant ($P \leq 0.005$) discrepancies between these libraries, but position 2 shows a strong bias. This analysis suggests that on the basis of single residue bias, the libraries could have been drawn from one another for all positions except position 2. As discussed in section 5.2.2, this single difference may indicate that polar residues contribute more strongly to dimerization of more dilute TMDs. However, the increased fraction of polar residues in winners from PGM-Low compared to PGM could result from the PGM-Low/200 $\mu\text{g/ml}$ CAM condition corresponding to a slightly higher stringency than PGM/400 $\mu\text{g/ml}$ CAM condition. The possibility that the stringency scales may not be perfectly matched between the two protocols is also consistent with PGM-low never exhibiting valine at position 5 at any selection stringency, but this observation lacks strong statistical significance.

I postulate that the reason discrete rules for TMD association have failed to be developed has in part been due to the insensitivity of the available assays. The strongest effects in my libraries (i.e. position 5 glycine) could be seen with either PGM or PGM-Low libraries. Additional amino acids are found in the top 1% PGM associating sequences versus the top 1% PGM-Low sequences. This is an indication that high expression assays are less stringent when determining TMD dimer strength. Nearly every assay available is based on the ToxR high expression promoter. For this reason these assays are lacking the stringency necessary to find the most important residues for TMD association.

Table 5.7 – Comparison of residue biases
between top 1% of PGM and PGM-Low clones.

	PGM from PGM-Low	PGM-Low from PGM
Position 1		
L	0.105	0.099
I	N/A	0.318
S	0.019	0.013
T	0.010	0.015
Position 2		
L	1.9×10^{-5}	3.8×10^{-5}
I	4.8×10^{-5}	1.1×10^{-3}
H	5.0×10^{-5}	2.3×10^{-5}
N	8.2×10^{-3}	3.2×10^{-4}
Position 3		
G	0.093	0.086
A	0.069	0.059
S	0.002	0.016
T	0.114	0.050
Position 4		
F	0.057	0.055
I	0.182	0.115
V	0.025	0.026
Position 5		
G	0.409	N/A*
A	0.411	N/A*
V	1.00	N/A*
Position 6		
A	0.081	0.078
V	0.106	0.100
I	0.225	0.224
T	0.114	0.050
Position 7		
G	0.104	0.097
A	0.103	0.087
S	0.004	0.007
T	0.044	0.013

*-N/A indicates that the calculation cannot be made since the residue in question is not found in PGM-Low library.

5.3.3 Combining the libraries and analyzing with a “pooled winners” rationale

I looked at the sequence variability in the winning clones (top 1%) from PGM and PGM-Low. Several sequences occurred multiple times in the winning pool. Duplicate sequences could result from biases in the parental library DNA, or they could result from differential growth characteristics, since our cells are grown for six hours after transformation but before plating to reach full expression of CAT. Biases in the parental DNA should be the same, since both libraries were cloned from the same PCR product. Since no strong correlation of duplicates was observed between libraries, I inferred that the duplicates arise primarily from differential growth characteristics, and only unique sequences were retained from each library in this analysis. These sequences can be found in Table 5.8. Given that many positions are not strongly biased, any pairs of residues that give significantly stronger dimerization should be observed with more than one residue at other positions, and so will be counted multiple times.

I used hypergeometric analysis to determine if PGM 400 and PGM-Low 200 could be considered as the same pool. Using a confidence limit of 0.005, there were no positions that displayed significant differences between these sets, as long as duplicates were excluded from the analysis. Therefore, I am confident that a set consisting of unique winners (refer to Table 5.8) could be analyzed further to obtain the strongest correlations that are shared between PGM 400 and PGM-Low 200. At the same time by combining these sets, I am accepting that any differences between PGM and PGM-Low winners will tend to cancel out and this nullifying effect will appear as non-specific contributions. To stress this point, I combine these sets in order to get at the most

significant single/pairwise biases, at the loss of minor single/pairwise biases. Statistical analysis was performed and the strongest P-values are listed in Table 5.9.

All the over-represented single site biases that I have reported for PGM 400 and PGM-Low 200 individually are retained in my combined winner analysis except for position 3, which becomes unbiased, see Figure 5.5. Under-represented biases are not in as good agreement, however. Biases against alanine are lost at positions 3 and 6, and against threonine at positions 1 and 7. Interestingly, the increased sample size increases the statistical significance of a modest bias against leucine at position 1 and 2 to the point that it can no longer be considered random. I infer that the observations are the very strongest single residue biases that can be drawn from PGM and PGM-Low.

Table 5.8 - Top 1% sequences combined

PGM	PGM-Low
1 LLSILLGVLLGILLS	1 LLSLLLAFLLGVLLA
2 LLTHLLSVLLGVLLG	2 LLSHLLAFLLGVLLS
4 LLLLLLAFLLGVLLG	3 LLLHLLGVLLGVLLG
5 LLLHLLGILLGVLLG	5 LLSILLGFLLGVLLG
6 LLTHLLAILLGALLA	6 LLSLLLAFLLGALLA
7 LLSHLLGFLLGALLS	7 LLLNLLGVLLGVLLG
8 LLLHLLGFLLGALLS	8 LLLNLLGILLGILLG
9 LLLNLLGVLLGVLLG	10 LLTHLLAFLLGTLA
10 LLLLLLVLLGVLLG	11 LLSHLLAFLLGALLA
11 LLSHLLAVLLGALLG	13 LLSHLLSVLLGALLA
17 LLSLLLGFLLGVLLG	14 LLLLLLGFLLGVLTT
20 LLSHLLAFLLGALLA	15 LLTHLLTFLLGALLT
21 LLSLLLAFLLGALLA	17 LLLNLLGFLLGVLLG
23 LLLILLGFLLGVLLG	18 LLSHLLTFLLGILLG
24 LLSLLLAFLLGVLLG	20 LLSLLLVLLGVLLS
25 LLTHLLTFLLGTLG	21 LLSHLLAVLLGVLLT
30 LLLILLGFLLGVLLS	22 LLLHLLGFLLGVLLG
33 LLLHLLGFLLGVLLG	25 LLTHLLAFLLGALLG
35 LLSLLLAFLLGVLLA	27 LLSHLLTFLLGVLLG
36 LLTLLLAFLLGVLLG	28 LLSHLLAFLLGTLA
37 LLTLLLGFLLGVLLG	30 LLSLLLAFLLGVLLG
40 LLLLLLGFLLGVLLA	31 LLSILLAFLLGVLLG
41 LLLHLLGVLLGILLG	34 LLSHLLGFLLGVLLG
42 LLSHLLAVLLGALLA	38 LLSHLLAFLLGVLLA
44 LLTLLLGFLLGALLT	41 LLSHLLAFLLGTLG
45 LLSHLLAVLLGILLA	42 LLSHLLAFLLGALLT
46 LLTHLLGVLLGVLLG	43 LLTLLLVLLGVLLS
47 LLSHLLTFLLGALLS	47 LLTHLLAVLLGVLLA
51 LLSHLLAVLLGALLT	50 LLLHLLGILLGVLLG
52 LLSLLLSFLLGALLA	51 LLSHLLAFLLGALLG
53 LLTHLLSFLLGALLS	52 LLSHLLTFLAVLLT
55 LLSLLLAFLLGALLS	56 LLSLLLAFLLGALLT
56 LLSILLGFLLGVLLG	57 LLLNLLGILLGVLLG
60 LLIHLLSFLLGTLLG	59 LLTHLLAFLLGTLG
61 LLSILLGVLLGVLLG	67 LLSLLLGFLLGVLTT
62 LLSHLLGFLLGVLLG	
64 LLLILLAFLLGVLLG	
66 LLSNLLGFLLGVLLG	
67 LLLLLLGFLLGVLLS	
71 LLLILLGVLLGVLLT	

Table 5.9 Single site biases for the combined PGM and PGM-Low libraries

Overrepresented Position		P-Value
5	G	3.6×10^{-30}
4	F	3.4×10^{-11}
2	H	2.1×10^{-9}
1	S	3.1×10^{-7}
6	V	9.3×10^{-4}
7	S	1.1×10^{-3}

Underrepresented Position		P-Value
5	A	9.7×10^{-16}
	V	7.5×10^{-9}
4	V	1.5×10^{-5}
	I	8.8×10^{-4}
3	T	2.2×10^{-4}
2	L	3.4×10^{-4}
7	A	2.4×10^{-3}
6	T	2.6×10^{-3}
1	L	4.2×10^{-3}

The top 1% PGM and PGM-Low unique clones were combined and compared to a combined reference set made up of unique, unselected PGM and PGM-Low clones.

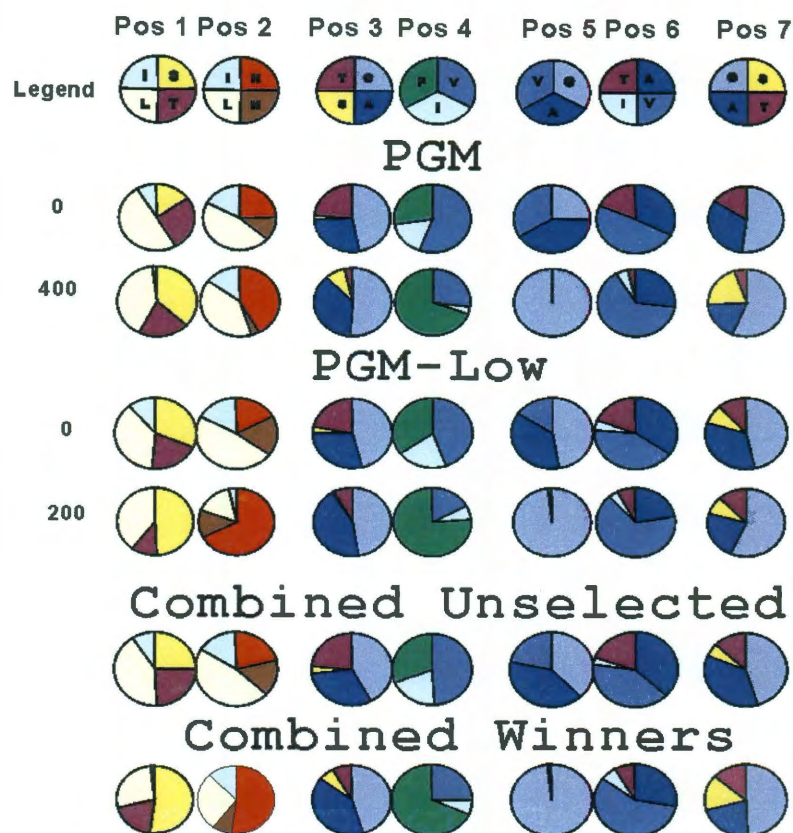


Figure 5.5 - PGM Library combination - Comparison of single site residue biases in PGM, PGM-Low, and the combined libraries (with unique sequences from PGM and PGM-Low) for unselected and top 1% clones. The excellent agreement between 'combined winners' and either parent set is a visual indication that the PGM and PGM-Low top 1% of associating sequences are drawn from the same pool. A statistical analysis confirms this conclusion (see text).

5.4 Comparing findings from PGM and PGM-Low to previous library studies

5.4.1 Statistical analysis of Russ et al's leulib data reveals that only the strongest bias is in agreement with PGM and PGM-low libraries

My library analyses can be most directly compared to the collection of strongly dimeric GpA like sequences identified by Russ et al using a TOXCAT-based library method and a poly-leucine background (Russ and Engelman, 2000). Like the work presented here, the authors' leulib data set contains variable residues at the spacing of the GpA interface, but there are differences in the permitted residues (Russ and Engelman, 2000) as described in sections 4.2.1 and 4.4.2.

To facilitate the comparison of my analysis to the previous study, I took the published leulib winning sequences (the most strongly dimerizing 0.001% of clones) and performed a hypergeometric analysis and calculated odds ratios based on the reported biases in the parental library (Table 5.10). Here I describe the similarities and differences I discovered between leulib and the PGM/PGM-Low winners, refer to Figures 5.3, 5.4 and 5.6.

The Leulib and PGM libraries are both strongly biased towards glycine at position 5 and neither alanine nor valine appears at this position in the winners of either library. In leulib, serine occurs relatively often at position 5, but the residue choices in PGM and PGM-Low only included glycine, alanine, and valine, so this is not directly comparable. In the leulib library, valine is over-represented at position 4 whereas my libraries show valine being biased against, but this is largely because phenylalanine, which is not present in leulib, is so strongly favored at PGM position 4. Position 3 was found to be

strongly biased for glycine in leulib, while the same position was unbiased or biased against glycine in PGM and PGM-Low, respectively. At position 2, leucine and valine are over-represented in leulib, whereas in the libraries I have presented histidine and asparagine are somewhat more common in the strongly dimerizing sequences. Position 6 is biased in favor of valine and against both threonine and alanine in PGM and PGM-Low, and leulib is biased in favor of both valine and alanine and against threonine although with only modest significance. Position 7 is highly biased for threonine in leulib, but in PGM and PGM-Low threonine is biased against or unbiased. Threonine is very under-represented at position 6 in PGM, but the lack of threonine in leulib is not statistically significant because it is not common in the unselected sequences. Leucine is over-represented at position 1 in leulib, whereas in the PGM libraries it is unbiased.

It is significant that the critical glycine identified in PGM is largely retained by leulib even though only one-third of the sequences encoded by PGM can occur in leulib. Differences we see between leulib and PGM type libraries likely stems from the different sequence contexts that can be supplied by each library.

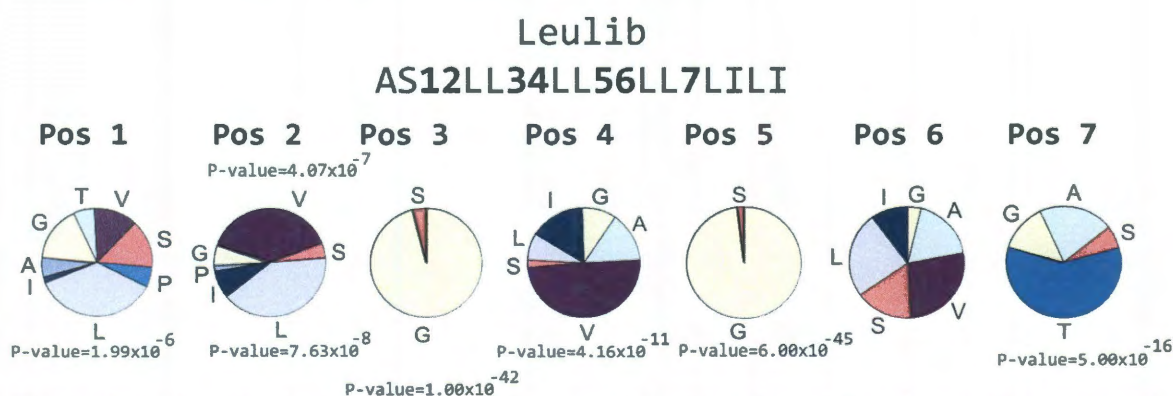


Figure 5.6 – Leulib single site residue frequencies with selected P-values – The strongest bias in Leulib is for G5, which is also the strongest bias in PGM libraries. The strong bias for T7 seen in Leulib is not seen in PGM winners even though this residue is present in the PGM library. Interpreting other bias differences is complicated by the differences in available residues in Leulib versus PGM libraries.

Table 5.10 Leulib Analysis.

A.Leulib Top 0.001% residue P-values.

Position 1	P	R	G	A	V	S	T	L	I	Total
Occurrence	3	0	8	3	6	7	3	18	1	49
P-Value	0.112	0.003	0.084	0.112	0.166	0.128	0.112	2.00×10^{-6}	0.019	
Position 2	P	R	G	A	V	S	T	L	I	Total
Occurrence	1	0	3	0	19	2	0	20	4	49
P-Value	0.019	0.003	0.112	0.003	4.07×10^{-7}	0.057	0.003	7.64×10^{-8}	0.161	
Position 3	P	R	G	A	V	S	T	L	I	Total
Occurrence	0	0	47	0	0	2	0	0	0	49
P-Value	0.003	0.003	1.00×10^{-42}	0.003	0.003	0.057	0.003	0.003	0.003	
Position 4	P	R	G	A	V	S	T	L	I	Total
Occurrence	0	0	5	7	24	1	0	4	8	49
P-Value	0.003	0.003	0.181	0.128	4.17×10^{-11}	0.019	0.003	0.161	0.084	
Position 5	P	R	G	A	V	S	T	L	I	Total
Occurrence	0	0	48	0	0	1	0	0	0	49
P-Value	0.003	0.003	6.00×10^{-45}	0.003	0.003	0.019	0.003	0.003	0.003	
Position 6	P	R	G	A	V	S	T	L	I	Total
Occurrence	0	0	2	9	13	8	0	12	5	49
P-Value	0.003	0.003	0.057	0.048	0.001	0.084	0.003	0.004	0.181	
Position 7	P	R	G	A	V	S	T	L	I	Total
Occurrence	0	0	7	10	0	3	29	0	0	49
P-Value	0.003	0.003	0.128	0.024	0.003	0.112	5.00×10^{-16}	0.003	0.003	

B.Leulib Top 0.001% residue Odds/Ratio.

Position 1	P	R	G	A	V	S	T	L	I
	0.55	0*	1.47	0.55	1.10	1.29	0.55	3.31	0.18
Position 2	P	R	G	A	V	S	T	L	I
	0.18	0*	0.55	0*	3.49	0.37	0*	3.67	0.73
Position 3	P	R	G	A	V	S	T	L	I
	0*	0*	8.63	0*	0*	0.37	0*	0*	0*
Position 4	P	R	G	A	V	S	T	L	I
	0*	0*	0.92	1.29	4.41	0.18	0*	0.73	1.47
Position 5	P	R	G	A	V	S	T	L	I
	0*	0*	8.82	0*	0*	0.18	0*	0*	0*
Position 6	P	R	G	A	V	S	T	L	I
	0*	0*	0.37	1.65	2.39	1.47	0*	2.20	0.92
Position 7	P	R	G	A	V	S	T	L	I
	0*	0*	1.29	1.84	0*	0.55	5.33	0*	0*

* A zero odds ratio is found for residues that do not occur at a given position.

5.4.2 Herrmann et al also identified a polar residue/glycine spacing in strong dimers

Langosch and colleagues used the low expression ToxR assay to investigate a large library that was based on a different spacing of variable positions in order to identify different patterns or motifs that contribute to TMD dimerization (Herrmann et al., 2009b). The heptad repeat used for their library design and the very large number of possible sequences they tried to sample limits the number of direct comparisons that can be made with my work, and they presented insufficient sequence data for me to perform a hypergeometric analysis and identify under- and over-represented residue trends, but one interesting comparison can be extracted. Their results stressed the importance of a histidine in the N-terminal part of the TMD in the most tightly associating sequences. Aligning this histidine with position 2 from PGM/PGM-Low shows that Herrmann's position 13, which was almost always a glycine, aligns with the glycine of PGM/PGM-Low position 5. It is noteworthy that the spacing of histidine and glycine imposed in my libraries by the parental GpA and BNIP3 TMDs recurs in the selected sequences of the Herrmann library, where any spacing of histidine and glycine were possible and where the intention of the design was to avoid the right-handed crossing of helices that had been identified previously.

Chapter 6 Pairwise Analysis and Combinatorial Effects

6.1 Defining the hypergeometric calculation for top winners

Extending the hypergeometric statistical approach from single-site residue propensities to pairwise correlations requires only slightly more complexity in the formulae that are applied, but because of the large number of pairwise possibilities, obtaining reliable statistics will require more sequences than for single-site trends. The strong similarities in single-site trends at most positions suggests that we could combine the sequenced clones from the two libraries to increase the sample size, and so increase the significance of any trend that is common to both libraries, but it should be noted that any trend that is specific to one library or the other would be decreased in significance by this approach. Although the hypergeometric analysis reveals differences at position 2 between the PGM and PGM-Low sequences obtained at the highest selection stringency, these differences are minor and the advantage gained by combinatorial analysis of the other sites outweighs this caveat. I combined the unique sequences from the two library data sets at the level of unselected clones, yielding 100 sequences from which parental single-site residue frequencies were calculated, and I combined the PGM 400 and PGM-Low 200 unique winners to give 75 total TMD sequences. With this larger data set, a reasonably robust analysis of pairwise correlations could be performed.

6.1.1 Data collection and restructuring the hypergeometric to uncover pairwise interactions

Pooling the isolates from the top 1% of PGM and PGM-Low libraries yields a set of tightly associating sequences that is similar to but distinct from either PGM or PGM-

Low. The hypergeometric analysis I describe calculates the likelihood of retrieving, by chance, the observed number of instances that two particular residues occur simultaneously in a set (here, the PGM/PGM-Low pooled winners) given the known incidence of each individual residue in an unselected set (from the pooled PGM/PGM-Low unselected sequences). This calculation requires four elements that are analogous to those used in the hypergeometric analysis described for single site propensities: k (occurrence), N (selection sample size), x (ratio of expected pairwise occurrence), and n (an expected ratio hypothetical sample size). These values are used in the hypergeometric equation, relisted below:

$$P(X = k) = \frac{\binom{m}{k} \binom{N-m}{n-k}}{\binom{N}{n}}$$

k is the # of successes in the sample
 n is the size of the sample
 m is the successes in the parent population
 N is the parent population size

The pooled selection sample size N can be found in Table 6.1a and 6.1b, for Unselected and Selected clones, respectively. The ‘ratio of expected pairwise occurrence’ is an estimate of the number of that particular combination that would be expected in a random distribution, which is calculated by multiplying the total number of occurrences that occur between two residue at different position and dividing by the total selection size. The actual occurrence of combinations k is obtained by counting the number of that particular pairwise combination that occurs in the library. The combinatorial possibilities are therefore derived from the seven degenerate positions that have been allowed in our small libraries. Each position varies between 3-4 possible choices (see Figure 4.4) which allows for 289 different pairwise combinations. The raw populations of these pairwise occurrences can be found in Table 6.2, and the hypergeometric equation is used to calculate the significances of the observed pairwise biases (see Table 6.3).

Table 6.1a - Pooled Unselected

Unselected Pool					
PGM 0- 1	LLTLLGVLLGALLG	PGM-Low	0- 1	LLSLLAVLLAVLLA	
PGM 0- 2	LLLHLLGVLLVALLA	PGM-Low	0- 2	LLSLLTFLLAALLA	
PGM 0- 3	LLTNLLTILLVTLT	PGM-Low	0- 3	LLIHLLGVLLGVLLG	
PGM 0- 4	LLLLLLGFLLVLLG	PGM-Low	0- 4	LLLLLLGFLLVLLG	
PGM 0- 5	LLTLLGVLLVLLG	PGM-Low	0- 5	LLLLLLGFLLVALLA	
PGM 0- 6	LLINLLGVLLGVLLG	PGM-Low	0- 6	LLLLLLGVLLGTLLS	
PGM 0- 8	LLTLLTILLATLLA	PGM-Low	0- 7	LLLLLLAFLLVALLG	
PGM 0- 9	LLIILLAVLLGVLLG	PGM-Low	0- 8	LLLLLLGFLLGVLLG	
PGM 0- 10	LLLLLLGVLLVALLG	PGM-Low	0- 9	LLIILLGVLLGALLT	
PGM 0- 11	LLLHLLGVLLGALLG	PGM-Low	0- 10	LLSLLGVLLGVLLG	
PGM 0- 12	LLSHLLAVLLAALLA	PGM-Low	0- 11	LLLLLLTVLLGALLA	
PGM 0- 13	LLLLLLSFLAVLLA	PGM-Low	0- 12	LLTNLLSILLGTLLS	
PGM 0- 14	LLTLLAFLLATLLG	PGM-Low	0- 13	LLSLLAILLAALLA	
PGM 0- 15	LLLLLLGVLLGVLLG	PGM-Low	0- 14	LLTNLLTILLATLLT	
PGM 0- 16	LLTNLLTILLATLLT	PGM-Low	0- 15	LLTHLLTILLGALLS	
PGM 0- 18	LLLLLLGVLLAVLLA	PGM-Low	0- 16	LLNLLGFLLVLLG	
PGM 0- 19	LLTHLLTILLATLLT	PGM-Low	0- 17	LLIHLLGVLLVLLG	
PGM 0- 23	LLSLLTFLLAALLA	PGM-Low	0- 18	LLLLLLGFLLAALLG	
PGM 0- 24	LLTLLGVLLAALLA	PGM-Low	0- 19	LLIILLGVLLGVLLG	
PGM 0- 25	LLLLLLAFLLVLLA	PGM-Low	0- 21	LLSLLAFLLVTLA	
PGM 0- 28	LLSHLLTVLLAALLA	PGM-Low	0- 22	LLIILLGVLLGVLLA	
PGM 0- 29	LLIILLGVLLGVLLG	PGM-Low	0- 23	LLSHLLAFLLGALLA	
PGM 0- 30	LLLLLLAFLLAALLA	PGM-Low	0- 24	LLTHLLTILLATLLG	
PGM 0- 31	LLTLLAVLLATLLT	PGM-Low	0- 26	LLNLLGVLLGVLLG	
PGM 0- 33	LLLLLLAFLLGVLLG	PGM-Low	0- 28	LLTNLLSVLLGILLS	
PGM 0- 36	LLTHLLTILLATLLA	PGM-Low	0- 29	LLLLLLAFLLVALLA	
PGM 0- 37	LLLLLLAVLLVLLG	PGM-Low	0- 31	LLIILLGVLLGVLLG	
PGM 0- 38	LLLLLLGVLLVLLG	PGM-Low	0- 32	LLSLLAVLLAALLS	
PGM 0- 40	LLIILLAFLLVLLG	PGM-Low	0- 33	LLTNLLTILLAVLLT	
PGM 0- 42	LLTILLGVLLAVLLG	PGM-Low	0- 34	LLINLLGVLLGVLLA	
PGM 0- 44	LLSLLAVLLAALLA	PGM-Low	0- 36	LLSLLAVLLAALLA	
PGM 0- 46	LLSNLLGVLLGVLLG	PGM-Low	0- 37	LLLLLLAFLLGVLLA	
PGM 0- 47	LLSHLLTVLLATLLT	PGM-Low	0- 38	LLLLLLGFLLAVLLA	
PGM 0- 48	LLLLLLAVLLVALLA	PGM-Low	0- 39	LLLLLLAFLLAALLG	
PGM 0- 50	LLLLLLAFLLGALLA	PGM-Low	0- 40	LLTHLLTILLATLLT	
PGM 0- 51	LLLLLLAVLLVLLA	PGM-Low	0- 42	LLSLLAVLLGALLA	
PGM 0- 52	LLIILLGFLLGVLLG	PGM-Low	0- 43	LLIILLGFLLGVLLG	
PGM 0- 53	LLIILLGVLLGALLG	PGM-Low	0- 44	LLSLLAILLATLLG	
PGM 0- 56	LLTHLLTVLLAALLT	PGM-Low	0- 45	LLSHLLTILLAILLG	
PGM 0- 57	LLTLLTVLLAALLG	PGM-Low	0- 47	LLSILLAVLLAALLG	
PGM 0- 59	LLIILLGFLLVLLG	PGM-Low	0- 48	LLSHLLTILLATLLS	
PGM 0- 60	LLTHLLTILLAALLA	PGM-Low	0- 49	LLSLLAFLLGALLG	
PGM 0- 61	LLSNLLAILLVALLA	PGM-Low	0- 51	LLTNLLGVLLGILLG	
PGM 0- 62	LLSLLGVLLGVLLG	PGM-Low	0- 52	LLLLLLTVLLAALLA	
PGM 0- 63	LLLHLLGVLLGVLLG	PGM-Low	0- 53	LLSHLLGFLLGTLLA	
PGM 0- 65	LLINLLGFLLVLLG	PGM-Low	0- 54	LLTHLLTILLAALLT	
		PGM-Low	0- 57	LLIILLAVLLGTLLA	
		PGM-Low	0- 58	LLTNLLGFLLGVLLG	
		PGM-Low	0- 60	LLSLLAFLLGALLA	
		PGM-Low	0- 66	LLTHLLTILLGTLLT	
		PGM-Low	0- 67	LLSLLTVLLAALLA	
		PGM-Low	0- 68	LLLLLLGVLLVLLG	
		PGM-Low	0- 69	LLTNLLAILLATLLT	
		PGM-Low	0- 72	LLIILLGFLLAVLLG	
n = 100					

Table 6.1b - Pooled Selected

Selected Pool				
PGM 400- 1	LLSILLGVLLGILLS	PGM-Low 200- 1	LLSLLLAFLLGVLLA	
PGM 400- 2	LLTHLLSVLLGVLLG	PGM-Low 200- 2	LLSHLLAFLLGVLLS	
PGM 400- 4	LLLLLLAFLLGVLLG	PGM-Low 200- 3	LLHLHLLGVLLGVLLG	
PGM 400- 5	LLHLHLLGILLGVLLG	PGM-Low 200- 5	LLSILLGFLLGVLLG	
PGM 400- 6	LLTHLLAILLGALLA	PGM-Low 200- 6	LLSLLLAFLLGALLA	
PGM 400- 7	LLSHLLGFLLGALLS	PGM-Low 200- 7	LLNLNLLGVLLGVLLG	
PGM 400- 8	LLHLHLLGFLLGALLS	PGM-Low 200- 8	LLNLNLLGILLGILLG	
PGM 400- 9	LLNLNLLGVLLGVLLG	PGM-Low 200- 10	LLTHLLAFLLGTLA	
PGM 400- 10	LLLLLLGVLLGVLLG	PGM-Low 200- 11	LLSHLLAFLLGALLA	
PGM 400- 11	LLSHLLAVLLGALLG	PGM-Low 200- 13	LLSHLLSVLLGALLA	
PGM 400- 17	LLSLLLGFLLGVLLG	PGM-Low 200- 14	LLLLLLGFLLGVLLT	
PGM 400- 20	LLSHLLAFLLGALLA	PGM-Low 200- 15	LLTHLLTFLLGALLT	
PGM 400- 21	LLSLLLAFLLGALLA	PGM-Low 200- 17	LLNLNLLGFLLGVLLG	
PGM 400- 23	LLLILLGFLLGVLLG	PGM-Low 200- 18	LLSHLLTFLLGILLG	
PGM 400- 24	LLSLLLAFLLGVLLG	PGM-Low 200- 20	LLSLLLVLLGVLLS	
PGM 400- 25	LLTHLLTFLLGTLG	PGM-Low 200- 21	LLSHLLAVLLGVLLT	
PGM 400- 30	LLLILLGFLLGVLLS	PGM-Low 200- 22	LLHLHLLGFLLGVLLG	
PGM 400- 33	LLHLHLLGFLLGVLLG	PGM-Low 200- 25	LLTHLLAFLLGALLG	
PGM 400- 35	LLSLLLAFLLGVLLA	PGM-Low 200- 27	LLSHLLTFLLGVLLG	
PGM 400- 36	LLTLLLAFLLGVLLG	PGM-Low 200- 28	LLSHLLAFLLGTLA	
PGM 400- 37	LLTLLLGFLLGVLLG	PGM-Low 200- 30	LLSLLLAFLLGVLLG	
PGM 400- 40	LLLLLLGFLLGVLLA	PGM-Low 200- 31	LLSILLAFLLGVLLG	
PGM 400- 41	LLHLHLLGVLLGILLG	PGM-Low 200- 34	LLSHLLGFLLGVLLG	
PGM 400- 42	LLSHLLAVLLGALLA	PGM-Low 200- 38	LLSHLLAFLLGVLLA	
PGM 400- 44	LLTLLLGFLLGALLT	PGM-Low 200- 41	LLSHLLAFLLGTLG	
PGM 400- 45	LLSHLLAVLLGILLA	PGM-Low 200- 42	LLSHLLAFLLGALLT	
PGM 400- 46	LLTHLLGVLLGVLLG	PGM-Low 200- 43	LLTLLLVLLGVLLS	
PGM 400- 47	LLSHLLTFLLGALLS	PGM-Low 200- 47	LLTHLLAVLLGVLLA	
PGM 400- 51	LLSHLLAVLLGALLT	PGM-Low 200- 50	LLHLHLLGILLGVLLG	
PGM 400- 52	LLSLLLSFLLGALLA	PGM-Low 200- 51	LLSHLLAFLLGALLG	
PGM 400- 53	LLTHLLSFLLGALLS	PGM-Low 200- 52	LLSHLLTFLLAVLLT	
PGM 400- 55	LLSLLLAFLLGALLS	PGM-Low 200- 56	LLSLLLAFLLGALLT	
PGM 400- 56	LLSILLGFLLGVLLG	PGM-Low 200- 57	LLNLNLLGILLGVLLG	
PGM 400- 60	LLIHLLSFLLGTLLG	PGM-Low 200- 59	LLTHLLAFLLGTLSS	
PGM 400- 61	LLSILLGVLLGVLLG	PGM-Low 200- 67	LLSLLLVLLGVLLT	
PGM 400- 62	LLSHLLGFLLGVLLG			
PGM 400- 64	LLLILLAFLLGVLLG			
PGM 400- 66	LLSNLLGFLLGVLLG			
PGM 400- 67	LLLLLLGFLLGVLLS			
PGM 400- 71	LLLILLGVLLGVLLT			

N = 75

Table 6.2 -Pooled selected pairwise occurrence.

	1S	1T	1L	1I		3G	3A	3S	3T
2H	21	10	7	1	4V	11	6	2	0
2N	1	0	5	0	4I	4	1	0	0
2L	12	4	5	0	4F	19	23	3	6
2I	5	0	4	0	5G	34	30	5	5
3G	11	4	19	0	5A	0	0	0	1
3A	22	6	2	0	5V	0	0	0	0
3S	2	2	0	1	6A	3	13	3	2
3T	4	2	0	0	6V	28	12	1	2
4V	9	4	6	0	6I	3	1	0	1
4I	0	1	4	0	6T	0	4	1	1
4F	30	9	11	1	7G	22	10	2	3
5G	38	14	21	1	7A	1	13	2	0
5A	1	0	0	0	7S	7	3	1	1
5V	0	0	0	0	7T	4	4	0	2
6A	15	5	1	0					
6V	19	6	18	0		4V	4I	4F	
6I	3	0	2	0	5G	19	5	50	
6T	2	3	0	1	5A	0	0	1	
7G	15	6	15	1	5V	0	0	0	
7A	12	3	1	0	6A	4	1	16	
7S	6	3	3	0	6V	12	3	28	
7T	6	2	2	0	6I	3	1	1	
					6T	0	0	6	
	2H	2N	2L	2I	7G	9	4	24	
3G	11	6	10	7	7A	4	1	11	
3A	18	0	10	2	7S	3	0	9	
3S	4	0	1	0	7T	3	0	7	
3T	6	0	0	0					
4V	11	2	3	3		5G	5A	5V	
4I	3	2	0	0	6A	21	0	0	
4F	25	2	18	6	6V	42	1	0	
5G	38	6	21	9	6I	5	0	0	
5A	1	0	0	0	6T	6	0	0	
5V	0	0	0	0	7G	37	0	0	
6A	15	0	6	0	7A	16	0	0	
6V	15	5	15	8	7S	12	0	0	
6I	3	1	0	1	7T	9	1	0	
6T	6	0	0	0					
7G	18	6	7	6		6A	6V	6I	6T
7A	10	0	6	0	7G	3	28	3	3
7S	6	0	4	2	7A	8	5	1	2
7T	5	0	1	1	7S	5	5	1	1
					7T	5	5	0	0

Table 6.3 - Statistical significance of pairwise biases

	<u>1S</u>	<u>1T</u>	<u>1L</u>	<u>1I</u>		<u>3G</u>	<u>3A</u>	<u>3S</u>	<u>3T</u>
2H	0.101	0.081	0.062	0.312	4V	0.091	0.137	0.228	0.215
2N	0.134	0.323	0.020	0.927	4I	0.115	0.271	0.718	0.670
2L	0.118	0.201	0.169	0.757	4F	0.061	0.080	0.224	0.110
2I	0.180	0.182	0.137	0.886	5G	0.092	0.094	0.182	0.168
3G	0.020	0.117	0.001	0.636	5A	0.636	0.670	0.942	0.070
3A	0.022	0.166	0.006	0.670	5V	1.000	1.000	1.000	1.000
3S	0.255	0.172	0.243	0.057	6A	0.008	0.035	0.113	0.268
3T	0.179	0.207	0.182	0.927	6V	0.009	0.041	0.160	0.190
4V	0.135	0.196	0.161	0.780	6I	0.205	0.271	0.718	0.270
4I	0.071	0.371	0.038	0.942	6T	0.062	0.127	0.270	0.300
4F	0.066	0.139	0.079	0.347	7G	0.038	0.046	0.262	0.229
5G	0.092	0.118	0.102	0.372	7A	0.004	0.006	0.198	0.275
5A	0.312	0.835	0.757	0.993	7S	0.126	0.151	0.363	0.371
5V	1.000	1.000	1.000	1.000	7T	0.195	0.202	0.514	0.144
6A	0.051	0.157	0.014	0.757					
6V	0.073	0.123	0.022	0.563		<u>4V</u>	<u>4I</u>	<u>4F</u>	
6I	0.223	0.391	0.245	0.942	5G	0.106	0.182	0.097	
6T	0.216	0.076	0.182	0.070	5A	0.780	0.942	0.347	
7G	0.059	0.157	0.039	0.303	5V	1.000	1.000	1.000	
7A	0.055	0.230	0.047	0.810	6A	0.166	0.348	0.099	
7S	0.167	0.203	0.225	0.854	6V	0.117	0.229	0.091	
7T	0.158	0.274	0.241	0.880	6I	0.095	0.240	0.109	
					6T	0.215	0.670	0.110	
	<u>2H</u>	<u>2N</u>	<u>2L</u>	<u>2I</u>	7G	0.140	0.132	0.095	
3G	0.020	0.035	0.133	0.062	7A	0.201	0.371	0.130	
3A	0.086	0.086	0.115	0.176	7S	0.229	0.448	0.134	
3S	0.143	0.670	0.348	0.546	7T	0.220	0.514	0.157	
3T	0.055	0.617	0.182	0.484					
4V	0.121	0.256	0.121	0.206		<u>5G</u>	<u>5A</u>	<u>5V</u>	
4I	0.223	0.053	0.243	0.546	6A	0.102	0.757	1.000	
4F	0.091	0.138	0.061	0.168	6V	0.092	0.326	1.000	
5G	0.092	0.168	0.102	0.141	6I	0.182	0.942	1.000	
5A	0.312	0.927	0.757	0.886	6T	0.168	0.927	1.000	
5V	1.000	1.000	1.000	1.000	7G	0.092	0.612	1.000	
6A	0.051	0.182	0.168	0.076	7A	0.112	0.810	1.000	
6V	0.018	0.131	0.076	0.071	7S	0.125	0.854	1.000	
6I	0.223	0.270	0.243	0.333	7T	0.135	0.113	1.000	
6T	0.055	0.617	0.182	0.484					
7G	0.102	0.047	0.077	0.129		<u>6A</u>	<u>6V</u>	<u>6I</u>	<u>6T</u>
7A	0.112	0.275	0.131	0.142	7G	0.004	0.022	0.217	0.229
7S	0.167	0.379	0.190	0.250	7A	0.044	0.051	0.371	0.230
7T	0.182	0.448	0.159	0.365	7S	0.126	0.133	0.363	0.371
					7T	0.087	0.173	0.514	0.448

6.1.2 Binning the effects into classes: over- and under-represented combinations

Our results show that all seven library positions take part in at least one significantly biased (P value ≤ 0.05) pairwise combination in the most strongly associating TMD dimers. Because a position 5 glycine occurred in every sequence, correlations with this position have no significance; instead, we must emphasize that the over- and under-represented combinations described here occur in a G5 background. The most significantly over- and under-represented pairwise combinations are listed with the constant glycine in Table 6.4. These residue triplets can be searched for in biological TMDs and compared to the results of other library experiments.

a. Over-represented trends

The most strongly over-represented pair seen from PGM/PGM-Low top 1% is G3V6 (glycine at position 3 and valine at position 6), with a statistical significance of $P = 0.009$ (see Table 6.5). A set of pairwise correlations that I term of ‘medium’ significance have P -values of ~ 0.02 and include H2V6, L1N2, S1G3, H2G3, and V6G7. After this level we see a grouping between 0.035-0.041. These are ‘low’ significance pairwise correlations consisting of N2G3, A3A6, L1I4, and A3V6. Finally a ‘weak’ set of combinations have been found: A3G7, N2G7, S1A6, and H2A6.

b. Under-represented trends

Combinations that were excluded from strongly associating dimers were less numerous and primarily involved alanines at position 3, 6, and 7. The ‘high’ significance correlations were G3A7, A6G7, A3A7, G3A6, and L1A6. Only one medium

significance combination, L1V6, was found in the under-represented set. The low significance set is more populated, L1V6, G3G7, and L1G7. Several weak combinations were found containing a position 7 alanine; A6A7, L1A7, and V6A7.

Table 6.4 - Significant Motifs

Overrepresented Motifs					
L1N2	LN..xx..Gx..x	H2G3	xH..Gx..Gx..x	A3G7	xx..Ax..Gx..G
S1G3	Sx..Gx..Gx..x	H2A6	xH..xx..GA..x	A3A6	xx..Ax..GA..x
N2G3	xN..Gx..Gx..x	H2V6	xH..xx..GV..x	A3V6	xx..Ax..GV..x
L1I4	Lx..xI..Gx..x	N2G7	xN..xx..Gx..G	V6G7	xx..xx..GV..G
S1A6	Sx..xx..GA..x	G3V6	xx..Gx..GV..x		
Underrepresented Motifs					
L1A6	Lx..xx..GA..x	G3A6	xx..Gx..GA..x	G3G7	xx..Gx..Gx..G
L1V6	Lx..xx..GV..x	V6A7	xx..xx..GV..A	G3A7	xx..Gx..Gx..A
L1A7	Lx..xx..Gx..A	A6G7	xx..xx..GA..G	A3A7	xx..Ax..Gx..A
L1G7	Lx..xx..Gx..G	A6A7	xx..xx..GA..A		

Table 6.5 - Significant Combinations

Overrepresented Set	P-Value	Underrepresented Set	P-Value
G3V6	0.009	G3A7	0.004
H2V6	0.018	A6G7	0.004
L1N2	0.020	A3A7	0.006
S1G3	0.020	G3A6	0.008
H2G3	0.020	L1A6	0.014
V6G7	0.022	L1V6	0.022
N2G3	0.035	G3G7	0.038
A3A6	0.035	L1G7	0.039
L1I4	0.038	A6A7	0.044
A3V6	0.041	L1A7	0.047
A3G7	0.046	V6A7	0.051
N2G7	0.047		
S1A6	0.051		
H2A6	0.051		

6.1.3 Interpretations of P-value magnitude

In the previous section (6.1.2;a and b) I present over- and under-represented pairwise combinations in rank order by P-value. I interpret the strongly over-represented and under-represented P-values to mean that these sets of residues either contribute or detract from dimerization strength, respectively. At these sample sizes, however, even the complete absence of a particular pairwise combination in our winners does not mean that this pair is unable to support dimerization – deeper sequencing might reveal another clone with that combination of residues. The demonstration that many different pairwise biases occur indicates that the effects of sequence context on the central glycine are very complex: many quite different combinations tend to be over-represented in the most associated TMDs. Our trends could be somewhat useful for prediction of the effects of mutations, but the complexity of our findings indicate that we will not be able to make perfect predictions for the effects of sequence changes on generic TMDs. Over-represented pairwise biases may fine tune TMD dimer stability of either a GpA-like or BNIP3-like interface, or they may lead to unique structures. More research will be needed to explore these hypotheses.

6.2 Comparing PGM/PGM-Low triplets to biological and library TMDs

Although an absolute register is not available to compare the PGM/PGM-Low combinatorial sets to biological TMDs, comparisons are still possible using relative positions if certain rules are followed. First, only residue triplets are directly comparable since all over- and under-represented combinatorial pairs obtained here contain the

background residue G5. Second, comparing relative positions assumes that the absolute depth of motifs within the bilayer does not affect the contribution of combination to dimerization. I think that this assumption is probably not true, but we have no way of accounting for how dimerization is influenced by depth, so any findings need to include this caveat. Here I compare my significantly biased pairwise combinations to combination that were also biased in the general TMD survey by Senes and colleagues (Senes et al., 2000). The combination GxxxGxxxG represents a tandem doublet of the TMD dimerization motif GxxxG, which has also been termed a ‘glycine zipper’ (Kim et al., 2005). In my work, I find the glycine zipper to be under-represented in tightly associating dimers (G3G7, $P < 0.038$) although several examples occur. In the survey of Senes *et al* this motif is found to be highly over-represented in biological TMDs in general. Interestingly, other potential variants on the glycine zipper, including GxxxGxxxA and AxxxGxxxA, are also under-represented in my winners, whereas the variant AxxxGxxxG, which is found in the tightly associating human BNIP3 TMD, is over-represented (A3G7, $P < 0.046$). The triplet GxxxGxxxA, which is strongly under-represented in my winners (G3A7, $P < 0.004$), occurs 10 times in the leulib of Russ et al (Russ and Engelman, 2000), showing that the availability of different residues at flanking positions can greatly alter the ranking of a ‘motif’ in a library. In the context of my library, the generalized glycine zippers described as GxxxGxxx(small) , (small)xxxGxxxG, and sometimes (small)xxx(small)xxx(small), where small is glycine alanine or serine, are generally under-represented (with the noted exception of A3G7). These findings show that library approaches could be used to further explore the contributions of glycine zippers, zipper variants, and their sequence context to TMD

dimerization. Although the limited variety of residues at the other degenerate positions in this library makes it hard to generalize our current findings to all zippers, it is clear that the generalized rules for zippers from other analyses do not apply in our PGM sequence context. This line of evidence again indicates that sequence context influences the ‘motif-driven’ self-association of TMDs enough to alter the rank ordering of the ‘motifs’.

I examined the PGM residue triplets to determine if any other previously identified motifs were over- or under-represented in the winners. Polar zipper motifs SxxSSxxT and SxxxSSxxT have been previously shown to drive TMD dimerization (Dawson et al., 2002), but these sequences are not allowed in my PGM type libraries. The pair SxxxS is allowed (as S1S3) but is unbiased in my winners. Strongly polar residues at position 2 (as occur in BNIP3) are over-represented in several pairwise combinations, but none of these involve a small polar residue at positions 1 or 3, which could be involved in hydrogen bonding at a BNIP3-like interface (Lawrie et al., 2010). Large polar residues are not significantly biased against in any pairwise PGM combinations. I therefore conclude that our library shows no evidence of bias towards or away from residues that would support the type of hydrogen bonding seen in the human BNIP3 TMD dimer structure (Sulistijo and MacKenzie, 2009). Thus, mechanisms for driving dimerization other than the BNIP3-type interface must be present in the library winners.

As noted above, the GxxxG motif occurs in both over- and under-represented motifs in the PGM winners. The over-represented pairs H2G3 ($P < 0.02$) or N2G3 ($P < 0.035$) combine with the invariant G5 to generate a GxxxG motif with the same spacing relative to a large polar residue. There is no occurrence of a glycine at position 3 in BNIP3 parental sequences, which use alanine or serine at position 3 in conjunction with a

strongly polar residue at position 2 and the G5G7 pair. Two of the H2G3 sequences have serines at position 7 (SHxxGFxxGAxxS and LHxxGFxxGAxxS), which has been shown to be incompatible with strong TOXCAT dimerization of human BNIP3 (SHxxAlxxGIxxG) (Lawrie et al., 2010). Shifting of the GxxxG from G5G7 to G3G5 may alter the dimer interface so that larger residues can be accommodated at position 7. How this new interface would utilize a strongly polar residue such as histidine is not clear: in one case, position 2 is a serine, which could be involved in BNIP3-like hydrogen bonding, but in the other position 2 is a leucine.

6.3 Conclusion/summary

Biased pairwise combinations were identified by examining the pooled top 1% of PGM and PGM-Low libraries. These biases vary in significance, which could be assessed more stringently by more extensive sequencing, but the findings in some cases support and in others go against motif or residue trends or themes currently thought to be involved in driving TMD dimerization. The most dramatic divergence from expectations is that the glycine zipper (GxxxGxxxG) and several variants are under-represented, although the AxxxGxxxG variant present in human BNIP3 is over-represented. The GxxxG sub-motif occurs in both over- and under-represented pairwise combinations.

Chapter 7 Conclusions and Discussion

7.1 Framing the discussion: caveats and counter-arguments

The analysis of PGM and PGM-Low TMD libraries has allowed me to identify novel sequence elements that contribute to the very tight association of BNIP3-like TMD dimers and also to probe contributions to less robust, but still significant, self-association of TMDs. I discuss the contributions to very strong interactions first, but it is important to note that the residue biases seen in modestly self-associating sequences have never been assessed previously by anyone, and so may represent the most novel of our findings. Finally, I discuss the results presented here as resulting from homodimerization, but any sequences in the PGM and PGM-Low libraries could form higher order structures, which would likely generate CAT expression and CAM resistance in TOXCAT, and any complexes could be stabilized by additional interactions with any of the hundreds of proteins that are inserted in the *E. coli* membrane. Thus, no particular sequence examined here has been directly shown to form dimers without other species. Although the TOXCAT method cannot discern between dimers and higher order species, and no TMD-TMD interaction library method can eliminate potential contributions from other species in the membrane, by comparing PGM-Low with PGM we decrease the chance of trimers and higher order species appearing in our analysis so that our trends can be interpreted as arising from dimers.

Highly associated PGM-Low clones show greatly decreased CAT expression compared to PGM clones, but both sets of winners consist of very similar, highly BNIP3-like TMDs, which are most likely to be dimeric given that the parental sequences all form dimers. The decreased CAT expression from PGM-Low is consistent with mass action:

diluting the membrane-inserted fusion protein reduces the fraction of dimer by decreasing the total amount of monomer. Importantly, the TMD-driven association of a higher order species would be even more greatly affected by dilution, since association depends on the fusion protein monomer concentration raised to the power of the oligomeric order. One possible way to interpret differences between PGM and PGM-Low is that the selection conditions of the latter greatly destabilizes higher order species relative to dimers, largely eliminating higher order oligomers from contributing to the observed sequences. The strong sequence similarities between PGM and PGM-Low across all selection strengths therefore indicate that higher order species make minimal contributions to our measures of residue biases that drive TMD homodimerization in PGM (and PGM-Low).

7.2 Lessons learned about sequence space from PGM/PGM-Low analysis

A ‘central glycine’ at position 5 of our library is critically required for strongest dimerization, and this glycine is very strongly over-represented even in sequences isolated at low stringency levels. I note that without sequencing every clone at a given stringency, we cannot state that this glycine is the only residue that supports robust dimerization for this library, but it is clearly the overwhelming favorite. This glycine is conserved between GpA, human BNIP3, and *C. elegans* BNIP3, and in the NMR structures of the first two TMD dimers this position represents the closest approach of the two helices, participates in backbone-backbone contacts, and contributes to dimer stability through non-canonical C α -H \cdots O=C hydrogen bonds (MacKenzie et al., 1997; Sulistijo and MacKenzie, 2009). It is tempting to propose that most sequences in our library that associate tightly do so through close helix-helix contacts, and likely make

non-canonical $\text{C}\alpha\text{-H}\cdots\text{O}=\text{C}$ hydrogen bonds (Senes et al., 2001): in this view, position 5 acts as a docking site that is primarily governed by steric forces and the geometrical limits of non-canonical hydrogen bonding. At intermediate stringencies our findings indicate that alanine can substitute at the central position, but valine cannot take the place of glycine. This appears to be a steric relationship and supports the idea that close approach of helices is the most important step to dimerization, and although glycine is the residue most often involved in donating a non-canonical hydrogen bond, other residues can do so, although the geometrical constraints are more severe (Senes et al., 2001), so the presence of an alanine does not strictly eliminate the possibility of intermonomer backbone-backbone hydrogen bonding. Because changing this critical glycine to alanine in GpA or BNIP3 abolishes dimerization, it seems likely that my library sequences containing alanine at position 5 do not use either a BNIP3-like or GpA-like interface. Although these clones appear largely at lower stringency, they could represent an interface completely distinct from those of the parental sequences.

The next most important position and residue that contribute to very strong TMD dimerization is a phenylalanine at position 4, three residues N-terminal to the central glycine. At this position, which is about one turn of the helix away from the central glycine, we permitted the apolar residues phenylalanine, isoleucine, and valine. Phenylalanine is preferred over the branched amino acid alternatives in tightly interacting TMDs, and since hydrophobicity remains relatively constant across these amino acids we attribute the ability for phenylalanine to its distinct shape, and possibly to aromatic-aromatic stacking or aromatic stacking on a $\text{C}\alpha\text{-H}$ of the central glycine. It may also be that in some cases, the phenylalanine side chain can swing away from the interface

around the side chain torsion angles χ_1 or χ_2 , to allow close approach of the helices, whereas the β -branched residues present bulky groups that cannot be rotated around χ_1 and so have no way of avoiding steric clashes imposed by backbone geometry.

Positions 3 and 7 are each one turn of helix away from the central glycine, and are assigned the same amino acid degeneracy in our libraries: glycine, alanine, serine, and threonine. Strong bias in the unselected sequences towards glycine and away from serine complicate the interpretation of these data, but the hypergeometric analysis reveals that glycine is not significantly over-represented at these positions in the winners. This is at first surprising since glycine at either of these two positions combines with the central glycine at position 5 to form the sequence GxxxG, a known dimerization motif. The lack of bias towards or against glycine at these positions indicates that glycine can support dimerization but that the other small residues, alanine, serine, and threonine, can as well. Alanines occur at either of these positions at the highest stringency, but whereas alanines are over-represented at position 3, they are under-represented and nearly absent at position 5. Pairwise analysis reveals that the glycine zipper GxxxGxxxG and the variants GxxxGxxxA and AxxxGxxxA are under-represented (although present) in the most tightly associating sequences, but the variant AxxxGxxxG is over-represented. Serines are strongly over-represented at these positions, but no pairwise combinations involving these serines is detected. There is no instance where a GpA-like combination involving a threonine is found.

These results show that the variants (small)xxxG and Gxxx(small) are not equivalent in the context of our library and that the tandem GxxxGxxxG motif does not give rise to stronger association than other combinations. Thus, whereas the central

glycine is absolutely required, many different combinations of small residues on the same face of the helix can give rise to strong dimerization. The over-represented AxxxGxxxG sequence indicates a bias towards this combination, perhaps through an interface similar to that of human BNIP3 which contains this motif. However, sequences such as GxxxGxxxA are not directly compatible with either the BNIP3 or GpA parent structures: mutations to match these residues disrupt the parents, and building these residues into the wild type structures causes clashes. It may be sequence changes in the flanking residues allow the interface to adjust slightly, and the structures of these library clones are subtle variations on the parents, but there could be significant differences in the geometry of the dimer interfaces relative to the parents.

Positions 1 and 2 were designed to test whether large polar residues are over-represented in tightly associating sequences and whether flanking small polar residues are correlated with the large polar residues. The results at position 2 were much more complex than anticipated: both strongly polar and hydrophobic residues occur at all levels of stringency, but both polar residues are significantly over-represented in the low expression library winners whereas both leucine and isoleucine are unbiased in the high expression library. Among the polar residues, histidine predominates at high stringency whereas asparagine predominates at low stringency in both libraries. The over-representation of strongly polar residues supports the idea of intermonomer side chain to side chain hydrogen bonding as has been seen in the structure of human BNIP3 (Sulistijo and MacKenzie, 2009), but the lack of a pairwise preference between large polar residues at position 2 and small hydrogen bonding residues at position 1 indicates that this type of interaction does not occur in all cases. Similarly, no pairwise correlation is detected

between large polar residues at position 2 and small hydrogen bonding residues at position 3, which is across the interface from position 2 and could serve as an alternate acceptor of hydrogen bonds.

At position 6, one residue after the central glycine, our analysis shows that valine is over-represented at the expense of threonine (and perhaps alanine), with isoleucine largely unbiased. This position is a valine in GpA and an isoleucine in human BNIP3, and although the detailed structures of these two show different crossing angles, these positions participate in making the ‘ridges’ that fit into ‘grooves’ of the opposite monomer, establishing the packing that is necessary to form a stable dimer. Three of the six most over-represented pairwise combinations involve a valine at position 6 (G3V6, H2V6, and V6G7), suggesting that having valine at this position favors the ability of the other positions to drive dimerization. Why the nearly isosteric threonine is much less able to do this is not clear.

I would like to stress that the single residue biases that I report here are simply independent observations that are statistically correlated with strong TMD association. While a single consensus sequence can be constructed from the strongest correlations, SHxxSFxxGIxxS, paying too much attention to the consensus ignores the richness and diversity of the assembled data. A library method can certainly be used to identify the single or few most tightly associating sequences, but our more extensive analysis has revealed trends that change with stringency and minimal biases that must reflect a great many different ways to drive dimerization. Although human BNIP3, GpA, and the consensus sequence described above are all members of this library, the diversity of the

sequences we have identified suggests that many members of the library differ from BNIP3 or GpA at least as much as these two systems differ from one another.

The prevalence of the central glycine at high stringency means that all pairs we find biases for in the PGM/PGM-Low data set occur in a (position 5 glycine) background and so correspond to triplets, not pairs. Only one of the twenty pairwise biases detected occurs between two adjacent residues, which suggest that residues adjacent in sequence may be slightly less likely than residues remote in sequence to show pairwise biases, since after excluding the central glycine just two of the fifteen possible paired positions are adjacent. Synergistic and antagonistic interactions between mutations at remote sites have been noted previously in quantitative (Doura and Fleming, 2004) and qualitative (Melnik et al., 2004) studies of interacting TMDs, and one proposed explanation is that the rigidity of the transmembrane helix enables packing changes at one position to alter the geometry at remote positions. Since every position in this library takes part in at least one biased pairwise interaction, I conclude that combinatorial effects should be the expected norm for interacting TMDs.

Many detailed conclusions can be drawn from our pairwise combinatorial analyses. First, large polar residues only appear to take part in over-represented pairs, but these are almost always are paired with non-polar residues. This rules out the likelihood that large side chain to small side chain hydrogen bonding provides a much stronger driving force for TMD dimerization than other basic mechanisms. The lack of under-represented pairwise correlations for large polar residues is less easy to interpret, but indicates that these residues do not block dimerization when paired with other residues in the library, perhaps because their side chains can be rotated to avoid clashes.

Hydrophobic residues such as leucine and valine take part in both over- and under-represented pairwise correlations. We find that GxxxG type dimeric motifs are not synergistic in tandem – although essentially all combinations of small residues are observed, only one instance of (small)xxxGxxx(small) over-representation is observed.

7.3 Additional questions in this library

The PGM/PGM-Low libraries were built with several restriction enzyme sites to permit the removal of certain residue choices before selection. Treatment of library DNA with AvrI will destroy sequences containing the central glycine, and the remaining subset of the library will serve as a future basis for researchers to answer several interesting questions: What are the tightest interacting sequences in this library that lack a glycine at position 5? Will single site trends be conserved without the central glycine? Are the combinatorial effects we report here independent of the central glycine? I expect that an entirely distinct set of sequence influences would be revealed in such an analysis, and that the ‘winners’ of this library will use very different interfaces to drive dimerization. It is also possible that higher order oligomers might be relatively common in this sequence space, which I would expect would be revealed by differences between clones isolated in the standard TOXCAT vector and the RBS-1 vector.

The idea that residues remote in the primary sequence are energetically coupled is suggested by over-represented pairwise correlations in PGM and PGM-Low library results. If these pairs are over-represented because of synergistic stabilizing interactions between the residues in question, then mutations at these positions should have predictable effects: single mutations at either site should be strongly disruptive, but

double mutations should be not much more disruptive than the single mutations. This lack of additivity could be tested using the TOXCAT assay in a quantitative mode (Duong et al., 2007) or biophysical methods such as ultracentrifugation. These two approaches have been taken to yield previous results about synergy and additivity in two systems (Doura and Fleming, 2004; Melnyk et al., 2004). Such experiments could directly establish synergy for any sequence tested, but because there are many instances of each over-represented pair, many mutants would need to be made and tested.

I expected that up to ninety unique sequences would be present in the top 1% of PGM and PGM-Low library, based on the size of the designed library and the plating experiments described in Chapter 5. Although I sequenced only 330 and 337 clones from these libraries, respectively, many duplicate sequences were encountered, and these redundancies confirm that the library is somewhat biased and does not contain identical levels of each possible clone. I actually obtained 40 (and 35) unique sequences in the top 1% of PGM (and PGM-Low), and although more extensive sequencing might identify another dozen clones, it appears that the “1 %” stringency levels established by my plating experiments actually contain only about half the expected number of unique sequences. This bias needs to be accounted for in analyzing the data (using the hypergeometric equation), but the bias is rather minimal in terms of how it affects the interpretation of the data.

7.4 Low expression constructs are well suited for combinatorial studies

Previous biological studies on highly dimeric TMDs have suffered from insensitivity of the assays to modestly disruptive mutations (Lawrie et al., 2010) even

though the same assay(s) work quantitatively for sequences with weaker interaction strengths (Duong et al., 2007). Using a TOXCAT vector variant that decreases expression of the ToxR-TMD-MBP fusion protein gives us similar rankings to standard TOXCAT, which allows us to conclude that our findings are largely derived from dimerization and not higher order interactions. My analysis reveals a more important role for strongly polar residues at position 2 in the PGM-Low library compared to the PGM library, which could be an indication that some of the sequences in PGM associate as higher order oligomers, but which I interpret as an indication that at lower concentrations of fusion protein the system discriminates better among different TMDs, thereby increasing the apparent sensitivity of the assay. For tightly associating systems, an optimized low expression assay has distinct advantages, although for a weakly associating system the signals could disappear into the noise.

The GpA TMD dimer associates tightly, but not so tightly that it saturates the standard TOXCAT assay: mutations that affect dimerization in detergents by just 1 kcal per mole have similar effects on the TOXCAT signal (Duong et al., 2007). In contrast, mutations that modestly disrupt dimerization of the human BNIP3 TMD in detergents have no effect on dimerization in TOXCAT: only strongly disruptive changes decrease the TOXCAT signal (Lawrie et al., 2010). The BNIP3 TOXCAT signal is about twice that of GpA, so any TMD that gives a similarly strong TOXCAT signal is probably a good candidate for analysis in the RBS1 low expression TOXCAT. General resistance to disruption is another indication that a low expression TOXCAT analysis should be employed. Another potential application for this system would be in heteromeric TMD interaction assays. Application of the low expression system alone or in combination

with the standard TOXCAT assay will undoubtedly produce valuable additional descriptions of the rich and complex hierarchy of single site and combinatorial residue contributions to TMD association.

7.5 The present and future of library investigations of TMD interactions

The design of low expression small library (<100,000) combinatorial experiments will provide researchers with a tool to assist in understanding the complex problem of TMD self association. The PGM/PGM-Low selection protocol described in this thesis is one of the first libraries specifically constructed to extract single site and pairwise combinatorial interactions in the residues that drive TMD dimerization. We see several instances where residues from parental TMDs are over-represented, and thus are inferred to contribute significantly to dimerization. This strongly supports the idea that these residues have been retained by nature by an evolutionary pressure to provide strong self association. Surprisingly, considerable freedom exists at the strongest selection level and allows sequences to diverge from known associating TMDs. The small library design is inherently more suited for combinatorial analysis. We determined many residue pairs that are over-represented or under-represented in strongly associating right handed TMD dimers. It is likely that there are several different kinds of dimerization mechanism occurring in PGM and PGM-Low. In order to generate general rules of dimerization more investigations will have to be undertaken to pull apart these mechanism by further library approaches. These approaches would intelligently limit the amino acid choices until clear sets of associating TMD types were identified. I suggest that in the future a library

scheme that excludes the central glycine will be used as a new tool aimed at finding and detailing intermediate TMD dimers.

Chapter 8 Methods

8.1 Library Design and Oligonucleotides

8.1.1 Design of transmembrane library sequences

The small library design (PGM/PGM-Low) was based on the idea that a commercially available medium length (~70 nucleotide) DNA oligonucleotide could be synthesized with degenerate codons at positions corresponding to interfacial residues and conserved codons at non-interfacial positions. In analyzing our library of ~10000 sequences, we wanted to be confident that we had only a small chance of missing any of the designed sequences, so we needed to generate clones numbering several times the total library size. Single pot PCR amplification allows me to generate a PCR insert that contains many orders of magnitude more molecules than the 40-50,000 needed to ensure the diversity of our library.

The short oligonucleotides used to PCR amplify the PGM insert add in-frame restriction enzyme sites compatible with TOXCAT plasmids. The hurdles in cloning a set of cells is covered in this chapter and include generating the proper size PGM insert and providing evidence PGM insert is competent for ligation.

8.1.2 PCR amplification of PGM sequences

Inherent bias is always a worry in combinatorial studies based on libraries. I worked with Sigma Aldrich (Woodlands, TX) to ensure that the primary degenerate oligo synthesis would contain as little bias as possible, i.e. that the mixtures of bases used at degenerate positions be as close to equimolar as possible. Sigma Aldrich ensured that a consistent mix of DNA bases reacted with the oligonucleotide in the proper ratio. By

keeping an open communication with Sigma Aldrich we then reduced the bias as much as possible since the actual synthesis of the oligonucleotide was not done in house.

Early attempts to produce a PGM library failed altogether because the prepared PGM inserts were not successfully ligated to cut TOXCAT plasmid, or because the efficiency of this process was unacceptably low given our goal of tens of thousands of unique clones. Exhaustive controls showed that the 75 bp PGM insert production was permissive at temperature throughout the standard PCR range (50-72°C). I took this as an indication that my reaction was specific and possibly contaminated. To alleviate these difficulties oligonucleotides were reordered with longer extensions (which would be chopped off by restriction digestion after purification).

By increasing the initial PGM insert PCR product size from ~75 bps to 101 bps I was able to overcome problems with manipulation of small DNAs. The 25 bp addition allows for better resolution of products in 2% agarose gels, and subsequent restriction enzyme digestion resulted in products of 79 bp and then 54 bp, both of which could be readily resolved on gels. This was highly preferred over the digestion of a 75 bp product that would be digested to ~65 bps; working with this system, it was difficult to tell whether the product was indeed cut. Extending the product size made it easy for me to discern when my products were properly cut.

8.1.3. Purification, restriction digests, and ligations

I made attempts to purify my products using either ethanol precipitation or commercial available kits, but these gave poor and unreliable yields. After several attempts at each, I worked with OmegaBioTek, the producer of my commercial

purification kit, and was able to optimize the system to recover 65-90% products. Three additional steps were critical to optimizing DNA recovery: first a isopropanol was added to 30%. Second, silica columns used to bind DNA were 'primed' by a soak and wash with 3M NaOH. Lastly elutions were done with 65°C sterile water instead of room temp water. Also a dry down step without heat was performed to guarantee removal of ethanol introduced during purification.

The PGM insert was generated by PCR using an extension/amplification scheme, see Figure 8.1 Insert generation. Product was amplified, purified, subjected to a double XbaI/DpnII digestion, and then again purified to reduce the losses associated with purification between single digestions. This product was stored at 4°C while TOXCAT vector was prepared using double digestion and phosphatase treatment. Ligations were set up in 20 µl reactions. Small quantities of unpurified ligations were transformed into 50 µl of 'high efficiency' DH5α cells. Such transformations typically yielded ~100-200 colonies using a standard protocol. Each colony represents one sequence from PGM library design.

8.1.4 Construction and transformation of PGM and PGM-Low required 'high' competent cell lines

a. Preparation of highly competent DH5α

As stated PGM library construction necessitated reaching a cloning level of 30000 sequences. Since I could make an accurate estimate of the number of possible attempts would be needed to create the library using (very expensive) commercially available cells, I took on preparation of competent cells in house. I used a publicly available

protocol (see Appendix Protocol 1 Competent Cell prep) to generate a ready supply of cells. My preparation was typically stored in aliquots of ~400 μ l. This gave two advantages: being able to use more cells if necessary & allowing me to make multiple transformations from the same stock tube. In this way a library could be built while maintaining the internal control that individual transformations were drawn from the same quality of cells.

b. Measuring competency

The level of competency usually obtained from in house preparation was on the order of 1×10^8 cfu (colony forming units) per μ g of vector in a 50 μ l reaction. This number is empirically determined by transforming a known amount of test plasmid (typically pUC19) into a given quantity of cells. I transformed 1 μ l DNA into 50 μ l prepared DH5 α cells. After completing standard heat shock transformation a small aliquot ranging from 1-10 μ l from the total volume of 1 ml was plated in triplicate on selective media. The average number of colonies from these plates was later used to calculate the total number of unique clones in the original transformation colonies. Negative (no plasmid) controls were also carried out to ensure that the competent cell stocks, buffers, and media were not contaminated.

c. Building PGM (by multiple transformation approach)

To over sample our 10,000 library sequences, multiple concurrent transformations were carried out. I sampled and plated a small amount of each transformation for transformation evaluations as described above. The bulk of each 1 ml transformation reaction was pooled with the others, and grown overnight at 37°C. An aliquot of the

overnight cultures was stored at -80°C as a glycerol stock, and the rest was mini-prepped to yield a single tube of plasmid containing many thousands of unique library sequences. An estimate of the number of unique sequences represented in the library was made using the plating aliquots described above; these are listed in Table 8.1 Library coverage. This approach was also used for cloning the PGM-Low library in the RBS-1 modified low expression TOXCAT vector.

Table 8.1 Library Coverage- From the total number of independent clones we can estimate the probability that any one sequence was not made, using a hypergeometric calculation. For PGM and PGM-Low we find that there < 5% and <3% chance, respectively, that any one sequence was missed.

PGM Library Construction

	Plates 10ul	Total Transformation Solution 301 ul
I	129	3882.9
II	308	9270.8
III	133	4003.3
IV	181	5448.1
V	152	4575.2
VI	222	6682.2
		<hr/>
		33862.5 Sum
		1125 Plated
		<hr/>
		32737.5 PGM Sequences

PGM-Low Library Construction

Transformation Set I

	Plates 10ul	Total Transformation Solution 301 ul
I	95	2859.5
II	82	2468.2
III	74	2227.4
IV	131	3943.1
V	160	4816
VI	116	3491.6
		<hr/>
		19805.8 Sum
		658 Plated
		<hr/>
		19147.8 Total Sequences

Transformation Set II

	Plates 10ul	Total Transformation Solution 301 ul
I	31	933.1
II	108	3250.8
III	109	3280.9
IV	56	1685.6
V	102	3070.2
VI	77	2317.7
VII	98	2949.8
VIII	83	2498.3
IX	49	1474.9
X	96	2889.6
		<hr/>
		24350.9 Sum
		809 Plated
		<hr/>
		23541.9 Total Sequences

DNA minipreps of Transformation Sets I & II were combined to make PGM-Low library.
 $19147.8 + 23541.9 = 42689.7$ PGM Low Sequences

Chapter 8.2 Plating Experiments

8.2.1 Preparation of electrocompetent NT326

The standard BioRad electrocompetent cell transformation protocol was performed on NT326 cells. This protocol can be found in Appendix Protocol 2 Electrocompetent Preparation. NT326 is an *E.coli* stock that is deficient in maltose binding protein (*malE*) and is the standard TOXCAT ready cell line.

8.2.2 Calibration of competency

The ability of cells to take up purified, isolated library DNA was calculated in the same way as for my prepared in house DH5 α cell line. Initial testing of transformed library DNA on selective media containing CAM resulted in little to no colonies. We alleviated this problem by giving cells a pregrowth period of approximately 7 hours without CAM, which gave the cells an acclimation period to begin producing CAT. This way, when cells were exposed to CAM on plates they were not put into a ‘shocked’ state by the effects of the antibiotic.

We find that colonies are capable of growing at 500 $\mu\text{g/ml}$ CAM for PGM and 200 $\mu\text{g/ml}$ CAM for PGM-Low. Empirically this matches well to Russ et al. previous GpA based TMDs library experiments with TOXCAT that used an identical cell line. Plating scales were empirically determined (see Table 8.2 Representary Plating Experiments). We found that in repeat experiments the same dilutions resulted in different concentration of cells. This would then result in a different quantity of cells plated. We therefore present plate experiment data that was done on the same day to reduce possible error.

Table 8.2 Representative Plating Experiments - Cells are transformed with library DNA, and grown without CAM. After an acclimation period of cells were plated in differ quantities and on varying levels of CAM. Colonies represent single TMD sequences and were pick for DNA analysis. PGM400 were collected without acclimation period and appear to make enough CAM to produce growth.

PGM

			Nt326 grown in 50ml lb 11:25am - 7:26pm (7hrs)					o.d.420 =0.5873 blanked against LB/Carb		plated 10ul		# colonies
										t1		240
										t2		323
										t3		357
												306.6667 Avg:
												32230.67 Total # col

volume plated ul	dilution factor vs prev	dilution procedure	cells/ml	cells in the plated aliquot		[CAM]				
						0	100	200	300	400
D0										
0	6	1ml in 6ml	7.97E+06	colonies in triplicate						
		6x (6x)		mean						
				s.d.						
D1										
50	25	0.2 in 5.0	318750.00	colonies in triplicate			750	195	22	
		25x (150x)		mean			494	115	3	
				s.d.			622	155	12.5	
D2										
50	25	0.2 in 5.0	12750.00	colonies in triplicate	476	407	81	26	6	
		25x (3750x)		mean	741	220	92	22	7	
				s.d.	608.5	313.5	86.5	24	6.5	
D3										
50	25	0.2 in 5.0	510.00	colonies in triplicate	31	18				
		25x (93750x)		mean	20	10				
				s.d.	25.5	14				

PGM-Low

			Nt326 grown in 50ml lb 12:57am - 9:30am 9hrs'					o.d.420 = 1.2247 blanked against LB/Carb		plated 10ul		# colonies
										t1		131
										t2		91
										t3		101
												107.6667 Avg:
												11208.1 Total # col

volume plated ul	dilution factor vs prev	dilution procedure	cells/ml	cells in the plated aliquot		[CAM]				
						0	100	200	300	400
D0										
0	6	1ml in 6ml	1.41E+07	colonies in triplicate						
		6x (6x)		mean						
				s.d.						
D1										
50	25	0.2 in 5.0	562500.00	colonies in triplicate			575	354	105	
		25x (150x)		mean			454	217	61	
				s.d.			514.5	285.5	83	
D2										
50	25	0.2 in 5.0	22500.00	colonies in triplicate	991	112	24	2	1	
		25x (3750x)		mean	901	160	15	1	0	
				s.d.	946	136	19.5	1.5	0.5	
D3										
50	25	0.2 in 5.0	900.00	colonies in triplicate	46	2				
		25x (93750x)		mean	44	2				
				s.d.	45	2				

8.2.3 Isolation of DNA

Selection plates that gave rise to colonies were picked and prepared using the Wizard SV Miniprep system, and each picked colony was also stored as a glycerol stock. To reduce cost and raise efficiency, I modified the protocol to recycle columns (HCl acid cleansing) and made several of the solutions in house. Original formulations resulted in very little to no sequence data, but this was remedied by reviewing Promega protocol releases. Using either the commercial materials or my own modified procedure, I get sequencing data that consistently contains 500 – 1000 readable bases. I have found that DNA concentration on gels to be unrelated to the amount of information I get back from our sequencing partner LoneStar Labs. Even so, I was able to build our library collection from overlapping sequencing attempts. In this way, if a sample failed to give usable information, it was repicked from glycerol stocks, grown, minipreped, and sent out as a new preparation. Protocol is summarized in Appendix Protocol 3 Wizard Miniprep. Archived copies of our raw sequencing data are available upon request.

8.2.4 Bias calculation

The identification of biases in the unselected library was carried out by sequencing clones from ampicillin-only plates. PGM and PGM-Low bias determinations were carried out independently and closely resemble each other (see Chapter 5 Table Bias Correlation). I therefore believe this bias originates primarily from the initial synthetic oligonucleotide.

8.3 Maltose complementation

8.3.1 Preparation of minimal media cultures

Carbon limited maltose media was prepared using standard techniques.

Ingredients found in Appendix Protocol 4 Minimal media were either autoclaved or filter sterilized. A small amount of Luria-Bertani medium (LB) was added to the mixture to act as a primer. Culture tubes were prepared using sterile technique and stored at 4°C.

Bacteria taken from rich media glycerol stocks and introduced into minimal media make use of the primer nutrients while modifying their metabolism to deal with limited resources. The primer is used up relatively quickly but gives the benefit of preparing minimal media cultures overnight, versus several days. Negative controls show that the rich nutrients from this primer are not carried to maltose plates.

8.3.2 Preparation of minimal media plates

Media plates were made using standard techniques. The recipe found in Appendix Protocol 4 Minimal media substitutes maltose instead of glucose for its limiting sugar. Aliquots (5 µl) of overnight minimal media cultures were plated in a dotting technique which allowed the testing of up to 20 specimens on a single plate.

References

- Adair, B.D., and Engelman, D.M. (1994). Glycophorin A helical transmembrane domains dimerize in phospholipid bilayers: a resonance energy transfer study. *Biochemistry* 33, 5539-5544.
- Adams, P.D., Engelman, D.M., and Brunger, A.T. (1996). Improved prediction for the structure of the dimeric transmembrane domain of glycophorin A obtained through global searching. *Proteins* 26, 257-261.
- Allen, S.J., Curran, A.R., Templer, R.H., Meijberg, W., and Booth, P.J. (2004). Folding kinetics of an alpha helical membrane protein in phospholipid bilayer vesicles. *J Mol Biol* 342, 1279-1291.
- Arkin, I.T., Adams, P.D., MacKenzie, K.R., Lemmon, M.A., Brunger, A.T., and Engelman, D.M. (1994). Structural organization of the pentameric transmembrane alpha-helices of phospholamban, a cardiac ion channel. *Embo J* 13, 4757-4764.
- Arkin, I.T., Brunger, A.T., and Engelman, D.M. (1997). Are there dominant membrane protein families with a given number of helices? *Proteins* 28, 465-466.
- Bocharov, E.V., Mineev, K.S., Volynsky, P.E., Ermolyuk, Y.S., Tkach, E.N., Sobol, A.G., Chupin, V.V., Kirpichnikov, M.P., Efremov, R.G., and Arseniev, A.S. (2008). Spatial structure of the dimeric transmembrane domain of the growth factor receptor ErbB2 presumably corresponding to the receptor active state. *J Biol Chem* 283, 6950-6956.
- Bocharov, E.V., Pustovalova, Y.E., Pavlov, K.V., Volynsky, P.E., Goncharuk, M.V., Ermolyuk, Y.S., Karpunin, D.V., Schulga, A.A., Kirpichnikov, M.P., Efremov, R.G., *et al.* (2007). Unique dimeric structure of BNip3 transmembrane domain suggests membrane permeabilization as a cell death trigger. *J Biol Chem* 282, 16256-16266.
- Bormann, B.J., Knowles, W.J., and Marchesi, V.T. (1989). Synthetic peptides mimic the assembly of transmembrane glycoproteins. *J Biol Chem* 264, 4033-4037.
- Braun, P., and von Heijne, G. (1999). The aromatic residues Trp and Phe have different effects on the positioning of a transmembrane helix in the microsomal membrane. *Biochemistry* 38, 9778-9782.
- Cady, S.D., Schmidt-Rohr, K., Wang, J., Soto, C.S., Degrado, W.F., and Hong, M. (2010). Structure of the amantadine binding site of influenza M2 proton channels in lipid bilayers. *Nature* 463, 689-692.

Choma, C., Gratkowski, H., Lear, J.D., and DeGrado, W.F. (2000). Asparagine-mediated self-association of a model transmembrane helix. *Nat Struct Biol* 7, 161-166.

Corver, J., Broer, R., van Kasteren, P., and Spaan, W. (2007). GxxxG motif of severe acute respiratory syndrome coronavirus spike glycoprotein transmembrane domain is not involved in trimerization and is not important for entry. *Journal of virology* 81, 8352-8355.

Cristian, L., Lear, J.D., and DeGrado, W.F. (2003). Determination of membrane protein stability via thermodynamic coupling of folding to thiol-disulfide interchange. *Protein Sci* 12, 1732-1740.

Dawson, J.P., Melnyk, R.A., Deber, C.M., and Engelman, D.M. (2003). Sequence context strongly modulates association of polar residues in transmembrane helices. *J Mol Biol* 331, 255-262.

Dawson, J.P., Weinger, J.S., and Engelman, D.M. (2002). Motifs of serine and threonine can drive association of transmembrane helices. *J Mol Biol* 316, 799-805.

Dews, I.C., and MacKenzie, K.R. (2007). Transmembrane domains of the syndecan family of growth factor coreceptors display a hierarchy of homotypic and heterotypic interactions. *Proc Natl Acad Sci U S A* 104, 20782-20787.

Dmitrova, M., Younes-Cauet, G., Oertel-Buchheit, P., Porte, D., Schnarr, M., and Granger-Schnarr, M. (1998). A new LexA-based genetic system for monitoring and analyzing protein heterodimerization in *Escherichia coli*. *Molecular & general genetics : MGG* 257, 205-212.

Doura, A.K., and Fleming, K.G. (2004). Complex interactions at the helix-helix interface stabilize the glycophorin A transmembrane dimer. *J Mol Biol* 343, 1487-1497.
Doura, A.K., Kobus, F.J., Dubrovsky, L., Hibbard, E., and Fleming, K.G. (2004). Sequence context modulates the stability of a GxxxG-mediated transmembrane helix-helix dimer. *J Mol Biol* 341, 991-998.

Duneau, J.P., Vegh, A.P., and Sturgis, J.N. (2007). A dimerization hierarchy in the transmembrane domains of the HER receptor family. *Biochemistry* 46, 2010-2019.

Duong, M.T., Jaszewski, T.M., Fleming, K.G., and MacKenzie, K.R. (2007). Changes in apparent free energy of helix-helix dimerization in a biological membrane due to point mutations. *J Mol Biol* 371, 422-434.

Engelman, D.M., Steitz, T.A., and Goldman, A. (1986). Identifying nonpolar transbilayer helices in amino acid sequences of membrane proteins. *Annu Rev Biophys Biophys Chem* 15, 321-353.

- Escher, C., Cymer, F., and Schneider, D. (2009). Two GxxxG-like motifs facilitate promiscuous interactions of the human ErbB transmembrane domains. *J Mol Biol* 389, 10-16.
- Faham, S., Yang, D., Bare, E., Yohannan, S., Whitelegge, J.P., and Bowie, J.U. (2004). Side-chain contributions to membrane protein structure and stability. *J Mol Biol* 335, 297-305.
- Fisher, L.E., Engelman, D.M., and Sturgis, J.N. (1999). Detergents modulate dimerization, but not helicity, of the glycophorin A transmembrane domain. *J Mol Biol* 293, 639-651.
- Fisher, L.E., Engelman, D.M., and Sturgis, J.N. (2003). Effect of detergents on the association of the glycophorin a transmembrane helix. *Biophys J* 85, 3097-3105.
- Fleishman, S.J., Schlessinger, J., and Ben-Tal, N. (2002). A putative molecular-activation switch in the transmembrane domain of erbB2. *Proc Natl Acad Sci U S A* 99, 15937-15940.
- Fleming, K.G. (2000). Probing stability of helical transmembrane proteins. *Methods Enzymol* 323, 63-77.
- Fleming, K.G. (2002). Standardizing the free energy change of transmembrane helix-helix interactions. *J Mol Biol* 323, 563-571.
- Fleming, K.G., Ackerman, A.L., and Engelman, D.M. (1997). The effect of point mutations on the free energy of transmembrane alpha-helix dimerization. *J Mol Biol* 272, 266-275.
- Freeman-Cook, L.L., Dixon, A.M., Frank, J.B., Xia, Y., Ely, L., Gerstein, M., Engelman, D.M., and DiMaio, D. (2004). Selection and characterization of small random transmembrane proteins that bind and activate the platelet-derived growth factor beta receptor. *J Mol Biol* 338, 907-920.
- Freeman-Cook, L.L., Edwards, A.P., Dixon, A.M., Yates, K.E., Ely, L., Engelman, D.M., and Dimaio, D. (2005). Specific locations of hydrophilic amino acids in constructed transmembrane ligands of the platelet-derived growth factor beta receptor. *J Mol Biol* 345, 907-921.
- Furthmayr, H. (1977). Structural analysis of a membrane glycoprotein: glycophorin A. *Journal of supramolecular structure* 7, 121-134.
- Gerber, D., Sal-Man, N., and Shai, Y. (2004). Two motifs within a transmembrane domain, one for homodimerization and the other for heterodimerization. *J Biol Chem* 279, 21177-21182.

Gratkowski, H., Lear, J.D., and DeGrado, W.F. (2001). Polar side chains drive the association of model transmembrane peptides. *Proc Natl Acad Sci U S A* 98, 880-885.

Gray, T.M., and Matthews, B.W. (1984). Intrahelical hydrogen bonding of serine, threonine and cysteine residues within alpha-helices and its relevance to membrane-bound proteins. *Journal of molecular biology* 175, 75-81.

Gurezka, R., and Langosch, D. (2001). In vitro selection of membrane-spanning leucine zipper protein-protein interaction motifs using POSSYCCAT. *J Biol Chem* 276, 45580-45587.

Herrmann, J.R., Fuchs, A., Panitz, J.C., Eckert, T., Unterreitmeier, S., Frishman, D., and Langosch, D. (2009a). Ionic interactions promote transmembrane helix-helix association depending on sequence context. *J Mol Biol* 396, 452-461.

Herrmann, J.R., Panitz, J.C., Unterreitmeier, S., Fuchs, A., Frishman, D., and Langosch, D. (2009b). Complex patterns of histidine, hydroxylated amino acids and the GxxxG motif mediate high-affinity transmembrane domain interactions. *J Mol Biol* 385, 912-923.

Hessa, T., Kim, H., Bihlmaier, K., Lundin, C., Boekel, J., Andersson, H., Nilsson, I., White, S.H., and von Heijne, G. (2005a). Recognition of transmembrane helices by the endoplasmic reticulum translocon. *Nature* 433, 377-381.

Hessa, T., Meindl-Beinker, N.M., Bernsel, A., Kim, H., Sato, Y., Lerch-Bader, M., Nilsson, I., White, S.H., and von Heijne, G. (2007). Molecular code for transmembrane-helix recognition by the Sec61 translocon. *Nature* 450, 1026-1030.

Hessa, T., White, S.H., and von Heijne, G. (2005b). Membrane insertion of a potassium-channel voltage sensor. *Science* 307, 1427.

Hirai, T., Subramaniam, S., and Lanyi, J.K. (2009). Structural snapshots of conformational changes in a seven-helix membrane protein: lessons from bacteriorhodopsin. *Curr Opin Struct Biol* 19, 433-439.

Jayasinghe, S., Hristova, K., and White, S.H. (2001). Energetics, stability, and prediction of transmembrane helices. *J Mol Biol* 312, 927-934.

Joh, N.H., Min, A., Faham, S., Whitelegge, J.P., Yang, D., Woods, V.L., and Bowie, J.U. (2008). Modest stabilization by most hydrogen-bonded side-chain interactions in membrane proteins. *Nature* 453, 1266-1270.

Killian, J.A., and von Heijne, G. (2000). How proteins adapt to a membrane-water interface. *Trends Biochem Sci* 25, 429-434.

- Kim, S., Chamberlain, A.K., and Bowie, J.U. (2004). Membrane channel structure of *Helicobacter pylori* vacuolating toxin: role of multiple GXXXG motifs in cylindrical channels. *Proc Natl Acad Sci U S A* *101*, 5988-5991.
- Kim, S., Jeon, T.J., Oberai, A., Yang, D., Schmidt, J.J., and Bowie, J.U. (2005). Transmembrane glycine zippers: physiological and pathological roles in membrane proteins. *Proc Natl Acad Sci U S A* *102*, 14278-14283.
- Kobus, F.J., and Fleming, K.G. (2005). The GxxxG-containing transmembrane domain of the CCK4 oncogene does not encode preferential self-interactions. *Biochemistry* *44*, 1464-1470.
- Kochendoerfer, G.G., Salom, D., Lear, J.D., Wilk-Orescan, R., Kent, S.B., and DeGrado, W.F. (1999). Total chemical synthesis of the integral membrane protein influenza A virus M2: role of its C-terminal domain in tetramer assembly. *Biochemistry* *38*, 11905-11913.
- Kyte, J., and Doolittle, R.F. (1982). A simple method for displaying the hydropathic character of a protein. *J Mol Biol* *157*, 105-132.
- Laage, R., Rohde, J., Brosig, B., and Langosch, D. (2000). A conserved membrane-spanning amino acid motif drives homomeric and supports heteromeric assembly of presynaptic SNARE proteins. *J Biol Chem* *275*, 17481-17487.
- Langosch, D., Brosig, B., Kolmar, H., and Fritz, H.J. (1996). Dimerisation of the glycophorin A transmembrane segment in membranes probed with the ToxR transcription activator. *J Mol Biol* *263*, 525-530.
- Lawrie, C.M., Sulistijo, E.S., and Mackenzie, K.R. (2010). Intermonomer Hydrogen Bonds Enhance GxxxG-Driven Dimerization of the BNIP3 Transmembrane Domain: Roles for Sequence Context in Helix-Helix Association in Membranes. *Journal of Molecular Biology* *396*, 924-936.
- Lear, J.D., Gratkowski, H., Adamian, L., Liang, J., and DeGrado, W.F. (2003). Position-dependence of stabilizing polar interactions of asparagine in transmembrane helical bundles. *Biochemistry* *42*, 6400-6407.
- Lemmon, M.A. (2009). Ligand-induced ErbB receptor dimerization. *Experimental cell research* *315*, 638-648.
- Lemmon, M.A., and Engelman, D.M. (1994). Specificity and promiscuity in membrane helix interactions. *FEBS Lett* *346*, 17-20.
- Lemmon, M.A., Flanagan, J.M., Hunt, J.F., Adair, B.D., Bormann, B.J., Dempsey, C.E., and Engelman, D.M. (1992a). Glycophorin A dimerization is driven by specific interactions between transmembrane alpha-helices. *J Biol Chem* *267*, 7683-7689.

- Lemmon, M.A., Flanagan, J.M., Treutlein, H.R., Zhang, J., and Engelman, D.M. (1992b). Sequence specificity in the dimerization of transmembrane alpha-helices. *Biochemistry* 31, 12719-12725.
- Lemmon, M.A., Treutlein, H.R., Adams, P.D., Brunger, A.T., and Engelman, D.M. (1994). A dimerization motif for transmembrane alpha-helices. *Nat Struct Biol* 1, 157-163.
- Li, E., and Hristova, K. (2006). Role of receptor tyrosine kinase transmembrane domains in cell signaling and human pathologies. *Biochemistry* 45, 6241-6251.
- Li, E., You, M., and Hristova, K. (2006). FGFR3 dimer stabilization due to a single amino acid pathogenic mutation. *J Mol Biol* 356, 600-612.
- Li, W., Metcalf, D.G., Gorelik, R., Li, R., Mitra, N., Nanda, V., Law, P.B., Lear, J.D., Degrado, W.F., and Bennett, J.S. (2005). A push-pull mechanism for regulating integrin function. *Proc Natl Acad Sci U S A* 102, 1424-1429.
- MacKenzie, K.R. (2006). Folding and Stability of alpha-Helical Integral Membrane Proteins. *Chem Rev* 106, 1931-1977.
- MacKenzie, K.R., and Engelman, D.M. (1998). Structure-based prediction of the stability of transmembrane helix-helix interactions: the sequence dependence of glycoporphin A dimerization. *Proc Natl Acad Sci U S A* 95, 3583-3590.
- MacKenzie, K.R., and Fleming, K.G. (2008). Association energetics of membrane spanning alpha-helices. *Curr Opin Struct Biol*.
- MacKenzie, K.R., Prestegard, J.H., and Engelman, D.M. (1997). A transmembrane helix dimer: structure and implications. *Science* 276, 131-133.
- Masi, A., Cicchi, R., Carloni, A., Pavone, F.S., and Arcangeli, A. (2010). Optical methods in the study of protein-protein interactions. *Adv Exp Med Biol* 674, 33-42.
- Melnyk, R.A., Kim, S., Curran, A.R., Engelman, D.M., Bowie, J.U., and Deber, C.M. (2004). The affinity of GXXXG motifs in transmembrane helix-helix interactions is modulated by long-range communication. *J Biol Chem* 279, 16591-16597.
- Mendrola, J.M., Berger, M.B., King, M.C., and Lemmon, M.A. (2002). The single transmembrane domains of ErbB receptors self-associate in cell membranes. *J Biol Chem* 277, 4704-4712.
- Merzlyakov, M., Chen, L., and Hristova, K. (2007). Studies of receptor tyrosine kinase transmembrane domain interactions: the EmEx-FRET method. *The Journal of membrane biology* 215, 93-103.

Merzlyakov, M., Li, E., Casas, R., and Hristova, K. (2006). Spectral Forster resonance energy transfer detection of protein interactions in surface-supported bilayers. *Langmuir* 22, 6986-6992.

Miller, V.L., Taylor, R.K., and Mekalanos, J.J. (1987). Cholera toxin transcriptional activator toxR is a transmembrane DNA binding protein. *Cell* 48, 271-279.

Miyauchi, K., Curran, R., Matthews, E., Komano, J., Hoshino, T., Engelman, D.M., and Matsuda, Z. (2006). Mutations of conserved glycine residues within the membrane-spanning domain of human immunodeficiency virus type 1 gp41 can inhibit membrane fusion and incorporation of Env onto virions. *Japanese journal of infectious diseases* 59, 77-84.

Oates, J., King, G., and Dixon, A.M. (2010). Strong oligomerization behavior of PDGFBeta receptor transmembrane domain and its regulation by the juxtamembrane regions. *Biochim Biophys Acta* 1798, 605-615.

Oxenoid, K., and Chou, J.J. (2005). The structure of phospholamban pentamer reveals a channel-like architecture in membranes. *Proc Natl Acad Sci U S A* 102, 10870-10875.

Petti, L.M., Irusta, P.M., and DiMaio, D. (1998). Oncogenic activation of the PDGF beta receptor by the transmembrane domain of p185neu*. *Oncogene* 16, 843-851.

Petti, L.M., Reddy, V., Smith, S.O., and DiMaio, D. (1997). Identification of amino acids in the transmembrane and juxtamembrane domains of the platelet-derived growth factor receptor required for productive interaction with the bovine papillomavirus E5 protein. *J Virol* 71, 7318-7327.

Popot, J.L., and Engelman, D.M. (1990). Membrane protein folding and oligomerization: the two-stage model. *Biochemistry* 29, 4031-4037.

Russ, W.P., and Engelman, D.M. (1999). TOXCAT: a measure of transmembrane helix association in a biological membrane. *Proc Natl Acad Sci U S A* 96, 863-868.

Russ, W.P., and Engelman, D.M. (2000). The GxxxG motif: a framework for transmembrane helix-helix association. *J Mol Biol* 296, 911-919.

Sapra, K.T., Balasubramanian, G.P., Labudde, D., Bowie, J.U., and Muller, D.J. (2008). Point mutations in membrane proteins reshape energy landscape and populate different unfolding pathways. *J Mol Biol* 376, 1076-1090.

Schneider, D., and Engelman, D.M. (2003). GALLEX, a measurement of heterologous association of transmembrane helices in a biological membrane. *J Biol Chem* 278, 3105-3111.

- Senes, A., Chadi, D.C., Law, P.B., Walters, R.F., Nanda, V., and DeGrado, W.F. (2007). E(z), a depth-dependent potential for assessing the energies of insertion of amino acid side-chains into membranes: derivation and applications to determining the orientation of transmembrane and interfacial helices. *J Mol Biol* 366, 436-448.
- Senes, A., Engel, D.E., and DeGrado, W.F. (2004). Folding of helical membrane proteins: the role of polar, GxxxG-like and proline motifs. *Curr Opin Struct Biol* 14, 465-479.
- Senes, A., Gerstein, M., and Engelman, D.M. (2000). Statistical analysis of amino acid patterns in transmembrane helices: the GxxxG motif occurs frequently and in association with beta-branched residues at neighboring positions. *J Mol Biol* 296, 921-936.
- Senes, A., Ubarretxena-Belandia, I., and Engelman, D.M. (2001). The Calpha ---H...O hydrogen bond: a determinant of stability and specificity in transmembrane helix interactions. *Proc Natl Acad Sci U S A* 98, 9056-9061.
- Seshadri, K., Garemyr, R., Wallin, E., von Heijne, G., and Elofsson, A. (1998). Architecture of beta-barrel membrane proteins: analysis of trimeric porins. *Protein Sci* 7, 2026-2032.
- Snider, C., Jayasinghe, S., Hristova, K., and White, S.H. (2009). MPEx: A tool for exploring membrane proteins. *Protein science*.
- Stanley, A.M., and Fleming, K.G. (2005). The transmembrane domains of ErbB receptors do not dimerize strongly in micelles. *J Mol Biol* 347, 759-772.
- Stanley, A.M., and Fleming, K.G. (2007). The role of a hydrogen bonding network in the transmembrane beta-barrel OMPLA. *J Mol Biol* 370, 912-924.
- Stites, W.E., Gittis, A.G., Lattman, E.E., and Shortle, D. (1991). In a staphylococcal nuclease mutant the side-chain of a lysine replacing valine 66 is fully buried in the hydrophobic core. *J Mol Biol* 221, 7-14.
- Stouffer, A.L., Acharya, R., Salom, D., Levine, A.S., Di Costanzo, L., Soto, C.S., Tereshko, V., Nanda, V., Stayrook, S., and DeGrado, W.F. (2008). Structural basis for the function and inhibition of an influenza virus proton channel. *Nature* 451, 596-599.
- Sulistijo, E.S., Jaszewski, T.M., and MacKenzie, K.R. (2003). Sequence-specific dimerization of the transmembrane domain of the "BH3-only" protein BNIP3 in membranes and detergent. *J Biol Chem* 278, 51950-51956.
- Sulistijo, E.S., and MacKenzie, K.R. (2006). Sequence Dependence of BNIP3 Transmembrane Domain Dimerization Implicates Side-chain Hydrogen Bonding and a Tandem GxxxG Motif in Specific Helix-Helix Interactions. *J Mol Biol* 364, 974-990.

- Sulistijo, E.S., and MacKenzie, K.R. (2009). Structural basis for dimerization of the BNIP3 transmembrane domain. *Biochemistry* 48, 5106-5120.
- Talbert-Slagle, K., and DiMaio, D. (2009). The bovine papillomavirus E5 protein and the PDGF beta receptor: it takes two to tango. *Virology* 384, 345-351.
- Tikhonova, I.G., and Costanzi, S. (2009). Unraveling the structure and function of G protein-coupled receptors through NMR spectroscopy. *Curr Pharm Des* 15, 4003-4016.
- Treutlein, H.R., Lemmon, M.A., Engelman, D.M., and Brunger, A.T. (1992). The glycophorin A transmembrane domain dimer: sequence-specific propensity for a right-handed supercoil of helices. *Biochemistry* 31, 12726-12732.
- White, S.H. (2003). Translocons, thermodynamics, and the folding of membrane proteins. *FEBS Lett* 555, 116-121.
- White, S.H. (2004). The progress of membrane protein structure determination. *Protein Sci* 13, 1948-1949.
- White, S.H. (2005). How hydrogen bonds shape membrane protein structure. *Adv Protein Chem* 72, 157-172.
- White, S.H., and von Heijne, G. (2005). Transmembrane helices before, during, and after insertion. *Curr Opin Struct Biol* 15, 378-386.
- White, S.H., and Wimley, W.C. (1999). Membrane protein folding and stability: physical principles. *Annu Rev Biophys Biomol Struct* 28, 319-365.
- Wimley, W.C. (2003). The versatile beta-barrel membrane protein. *Curr Opin Struct Biol* 13, 404-411.
- Yau, W.M., Wimley, W.C., Gawrisch, K., and White, S.H. (1998). The preference of tryptophan for membrane interfaces. *Biochemistry* 37, 14713-14718.
- Yin, H., Slusky, J.S., Berger, B.W., Walters, R.S., Vilaire, G., Litvinov, R.I., Lear, J.D., Caputo, G.A., Bennett, J.S., and DeGrado, W.F. (2007). Computational design of peptides that target transmembrane helices. *Science* 315, 1817-1822.
- Yohannan, S., Faham, S., Yang, D., Grosfeld, D., Chamberlain, A.K., and Bowie, J.U. (2004a). A C alpha-H...O hydrogen bond in a membrane protein is not stabilizing. *J Am Chem Soc* 126, 2284-2285.
- Yohannan, S., Yang, D., Faham, S., Boulting, G., Whitelegge, J., and Bowie, J.U. (2004b). Proline substitutions are not easily accommodated in a membrane protein. *J Mol Biol* 341, 1-6.

You, M., Li, E., Wimley, W.C., and Hristova, K. (2005). Forster resonance energy transfer in liposomes: measurements of transmembrane helix dimerization in the native bilayer environment. *Anal Biochem* 340, 154-164.

Zhou, F.X., Cocco, M.J., Russ, W.P., Brunger, A.T., and Engelman, D.M. (2000). Interhelical hydrogen bonding drives strong interactions in membrane proteins. *Nat Struct Biol* 7, 154-160.

Zhou, F.X., Merianos, H.J., Brunger, A.T., and Engelman, D.M. (2001). Polar residues drive association of polyleucine transmembrane helices. *Proc Natl Acad Sci U S A* 98, 2250-2255.

Appendices

Appendix Table 1.1A PGM Total & Total Unique

[illegible]

PGM - Low

[illegible]

Appendix Table 1.2A - Significance Analysis

PGM - Library P-value by selection level.

Pos 1	CAM (µg/ml)	L	I	S	T
	100	5.4×10^{-3}	9.4×10^{-3}	2.2×10^{-3}	4.3×10^{-2}
	200	7.3×10^{-2}	4.9×10^{-2}	5.9×10^{-3}	6.4×10^{-3}
	300	4.0×10^{-2}	1.0×10^{-2}	5.6×10^{-5}	7.2×10^{-2}
	400	4.5×10^{-2}	2.5×10^{-2}	3.5×10^{-3}	7.2×10^{-2}
Pos 2		L	I	H	N
	100	1.8×10^{-2}	1.2×10^{-1}	2.9×10^{-2}	3.4×10^{-5}
	200	1.2×10^{-2}	1.2×10^{-1}	2.3×10^{-2}	4.9×10^{-6}
	300	8.5×10^{-2}	1.2×10^{-1}	9.6×10^{-2}	2.2×10^{-2}
	400	6.2×10^{-2}	1.3×10^{-1}	1.2×10^{-3}	1.5×10^{-2}
Pos 3		G	A	S	T
	100	2.4×10^{-4}	3.9×10^{-3}	2.5×10^{-2}	8.6×10^{-3}
	200	5.7×10^{-7}	2.0×10^{-2}	9.2×10^{-2}	5.6×10^{-5}
	300	8.9×10^{-6}	2.3×10^{-2}	8.9×10^{-2}	2.9×10^{-5}
	400	8.3×10^{-2}	2.9×10^{-2}	2.9×10^{-3}	1.4×10^{-5}
Pos 4		F	I	V	
	100	8.0×10^{-2}	8.5×10^{-2}	9.6×10^{-2}	
	200	4.1×10^{-2}	1.5×10^{-2}	9.1×10^{-2}	
	300	1.4×10^{-1}	1.5×10^{-3}	9.1×10^{-6}	
	400	1.9×10^{-1}	1.0×10^{-3}	8.0×10^{-6}	
Pos 5		G	A	V	
	100	6.0×10^{-27}	6.1×10^{-7}	1.2×10^{-13}	
	200	9.4×10^{-41}	6.5×10^{-15}	5.2×10^{-14}	
	300	3.6×10^{-40}	1.1×10^{-14}	8.0×10^{-14}	
	400	2.0×10^{-35}	1.0×10^{-13}	1.3×10^{-11}	
Pos 6		A	V	I	T
	100	7.0×10^{-3}	6.1×10^{-4}	1.6×10^{-3}	8.6×10^{-5}
	200	9.2×10^{-3}	3.1×10^{-4}	6.9×10^{-3}	3.6×10^{-5}
	300	5.3×10^{-4}	3.5×10^{-6}	6.8×10^{-2}	2.3×10^{-4}
	400	7.5×10^{-2}	4.8×10^{-3}	1.2×10^{-1}	3.7×10^{-4}
Pos 7		G	A	S	T
	100	1.7×10^{-2}	2.1×10^{-2}	2.2×10^{-1}	6.3×10^{-2}
	200	1.9×10^{-3}	1.5×10^{-4}	2.2×10^{-1}	1.9×10^{-3}
	300	1.4×10^{-2}	1.3×10^{-3}	1.8×10^{-5}	1.2×10^{-3}
	400	8.5×10^{-2}	6.9×10^{-3}	2.3×10^{-5}	9.5×10^{-3}

PGM Library Odds Ratios.

Pos 1	CAM (µg/ml)	L	I	S	T
	100	0.71	1.99	1.84	0.72
	200	0.92	2.09	1.97	0.21
	300	0.85	1.96	1.72	0.54
	400	0.83	0.19	2.36	0.77
Pos 2		L	I	H	N
	100	0.77	1.08	0.65	2.49
	200	0.75	1.05	0.63	2.65
	300	0.94	0.90	0.88	1.63
	400	0.86	0.87	1.73	0.28
Pos 3		G	A	S	T
	100	1.44	0.52	3.31	0.53
	200	1.61	0.65	2.42	0.12
	300	1.54	0.66	2.45	0.23
	400	1.09	1.35	4.92	0.14
Pos 4		F	I	V	
	100	1.14	0.75	1.01	
	200	1.26	0.48	1.03	
	300	1.58	0.16	0.97	
	400	2.52	0.20	0.49	
Pos 5		G	A	V	
	100	3.37	0.32	0.00	
	200	3.81	0.04	0.00	
	300	3.81	0.04	0.00	
	400	3.87	0.00	0.00	
Pos 6		A	V	I	T
	100	0.61	1.39	*1.73	0.45
	200	0.64	1.41	*3.36	0.15
	300	0.47	1.55	*2.27	0.22
	400	0.83	1.33	*2.05	0.18
Pos 7		G	A	S	T
	100	1.22	0.70	*0.86	0.64
	200	1.32	0.42	*2.80	0.27
	300	1.23	0.52	*0.57	1.09
	400	1.08	0.57	*4.10	0.33

* Position 6I & 7S did not occur in PGM Unselected sequences.
Here p-values & odds ratio are calculated against combined
PGM/PGM-Low unselected populations.

Appendix Table 1.2B Significance Analysis

PGM-Low - Library P-value by selection level.

Pos 1	CAM (µg/ml)	L	I	S	T
50		4.6×10^{-2}	1.0×10^{-3}	4.6×10^{-3}	5.0×10^{-4}
100		4.6×10^{-3}	1.5×10^{-1}	1.8×10^{-2}	1.9×10^{-1}
150		5.0×10^{-2}	1.2×10^{-4}	1.5×10^{-3}	1.3×10^{-2}
200		9.3×10^{-2}	3.6×10^{-4}	1.1×10^{-3}	2.5×10^{-2}
Pos 2		L	I	H	N
50		6.8×10^{-7}	1.3×10^{-4}	1.6×10^{-3}	1.2×10^{-11}
100		1.9×10^{-8}	2.0×10^{-3}	2.5×10^{-3}	3.9×10^{-4}
150		5.8×10^{-8}	6.1×10^{-5}	3.9×10^{-8}	2.0×10^{-2}
200		6.4×10^{-8}	2.5×10^{-4}	2.9×10^{-9}	9.5×10^{-2}
Pos 3		G	A	S	T
50		6.0×10^{-7}	5.8×10^{-3}	2.4×10^{-1}	3.4×10^{-7}
100		2.7×10^{-5}	1.7×10^{-2}	2.7×10^{-1}	2.7×10^{-5}
150		1.6×10^{-2}	9.3×10^{-3}	1.3×10^{-1}	1.6×10^{-4}
200		9.4×10^{-2}	3.6×10^{-3}	2.5×10^{-1}	8.7×10^{-4}
Pos 4		F	I	V	
50		5.0×10^{-2}	1.6×10^{-2}	2.3×10^{-3}	
100		3.2×10^{-2}	2.7×10^{-3}	2.7×10^{-2}	
150		3.4×10^{-2}	1.1×10^{-3}	6.7×10^{-6}	
200		4.9×10^{-2}	6.4×10^{-4}	3.1×10^{-6}	
Pos 5		G	A	V	
50		1.6×10^{-26}	8.6×10^{-15}	4.5×10^{-6}	
100		6.0×10^{-19}	5.4×10^{-13}	3.8×10^{-4}	
150		1.6×10^{-20}	2.1×10^{-12}	1.8×10^{-6}	
200		1.8×10^{-2}	2.1×10^{-12}	9.0×10^{-6}	
Pos 6		A	V	I	T
50		1.5×10^{-5}	4.0×10^{-10}	2.2×10^{-1}	8.9×10^{-5}
100		6.5×10^{-7}	1.2×10^{-12}	1.9×10^{-1}	3.4×10^{-5}
150		2.1×10^{-3}	3.1×10^{-8}	2.2×10^{-1}	1.5×10^{-4}
200		9.5×10^{-3}	3.0×10^{-5}	2.3×10^{-1}	4.9×10^{-3}
Pos 7		G	A	S	T
50		4.7×10^{-5}	1.2×10^{-5}	9.8×10^{-2}	1.0×10^{-1}
100		3.1×10^{-4}	9.3×10^{-6}	9.4×10^{-1}	1.0×10^{-1}
150		7.0×10^{-4}	9.0×10^{-2}	1.7×10^{-1}	1.4×10^{-1}
200		2.1×10^{-2}	1.7×10^{-2}	1.7×10^{-1}	1.5×10^{-1}

PGM-Low Library Odds Ratios.

Pos 1	CAM (µg/ml)	L	I	S	T
50		1.18	2.15	0.58	0.67
100		1.82	0.88	0.53	0.22
150		1.39	0.70	0.64	0.98
200		1.06	0.00	1.55	0.55
Pos 2		L	I	H	N
50		0.41	0.16	1.53	2.90
100		0.26	0.32	1.75	2.94
150		0.33	0.09	2.68	2.06
200		0.34	0.17	3.85	0.77
Pos 3		G	A	S	T
50		1.62	0.79	0.44	0.06
100		1.78	0.49	0.88	0.06
150		1.56	0.88	0.00	0.14
200		1.04	1.51	0.47	0.34
Pos 4		F	I	V	
50		0.81	0.55	1.36	
100		0.76	0.27	1.53	
150		1.25	0.30	1.13	
200		2.18	0.29	0.40	
Pos 5		G	A	V	
50		2.10	0.00	0.00	
100		2.10	0.00	0.00	
150		2.07	0.00	0.10	
200		2.07	0.04	0.00	
Pos 6		A	V	I	T
50		0.36	1.88	0.89	0.30
100		0.28	1.99	1.17	0.15
150		0.54	1.82	0.66	0.16
200		0.64	1.59	0.94	0.39
Pos 7		G	A	S	T
50		1.50	0.34	0.59	1.27
100		1.51	0.33	0.58	1.25
150		1.43	0.38	0.98	1.13
200		1.23	0.67	0.94	1.07

Appendix Protocol 1 Competent Cell Preparation

1. DH5 α cells are grown in 1 liter of LB at 20°C to an OD₆₀₀ of ~ 0.3.
2. Cells are pelleted at 3000g, 4°C and resuspended in 80 ml of ice cold CCMB80 buffer.
3. Samples are stored on ice for 20 minutes.
4. Cells are pelleted at 3000g, 4°C and resuspended in 10 ml of ice cold CCMB80 buffer.
5. Resuspended cells are mixed with SOC (50 μ l:200 μ l). OD₆₀₀ is adjusted to 1.0-1.5 by addition of CCMB80
6. Cells are aliquoted and stored at -80°C indefinitely.

CCMB80 buffer

10 mM KOAc pH 7.0 (10 ml of a 1M stock/L)

80 mM CaCl₂.2H₂O (11.8 g/L)

20 mM MnCl₂.4H₂O (4.0 g/L)

10 mM MgCl₂.6H₂O (2.0 g/L)

10% glycerol (100 ml/L)

Adjust pH DOWN to 6.4 with 0.1N HCl if necessary

Preparing highly competent cells relies on use of these specific buffers, and growth in low temperature. Glassware was half filled with sterile water and autoclaved. Water was poured off, and glassware was re-autoclaved. This step is essential for removal of detergents.

Adapted from http://openwetware.org/wiki/TOP10_chemically_competent_cells

Appendix Protocol 2 Electrocompetent Cell Preparation

1. Pick NT326 master stock into large scale growth (1/2 liter).
2. Cultures are raised to approx. 0.5-0.7 OD₆₀₀. Once cells reach target OD, cells are chilled then centrifuged at 4000 xg for 15 minutes.
3. Resulting pellet is resuspended then subjected to repetitive wash and centrifuge cycles of increasing amounts of ice-cold 10% glycerol.

Wash 1: 500 ml 10% glycerol

Wash 2: 250 ml 10% glycerol

Wash 3: ~20 ml 10% glycerol

4. After final decantation, cell are resuspended in 2ml 10% glycerol and aliquoted for storage at -80°C.

Protocol was adapted from MicroPulser Electroporation Apparatus Operating Instructions and Applications Guide by Biorad Corporation.

Appendix Protocol 3 Wizard Miniprep

1. Cells from overnight culture are collected in the centrifuged briefly at top speed.
2. Cells are resuspended and lysed step wise in Resuspension and Lysis Buffer.
Resuspension solution consisted of a mixture of RNase, Tris buffering agent, and EDTA.
3. The addition of 10 μ l alkaline protease is optional. After approximately five minutes the digestion is quenched by Neutralization Solution with gentle mixing by inversion.
4. Mixture is separated by benchtop centrifugation (14000g) for 10 minutes.
5. Supernatant is bound to silica spin columns. Column are washed with ethanol/ Tris/HCl guanidinium which removes any contaminants while keeping DNA adhered to silica beads.
6. DNA is eluted with sterile water.

*- Protocol was adapted and interpreted from Wizard Plus SV Minipreps DNA Purification System produced by Promega Corporation. Resuspension Buffer, Lysis Buffer, and Neutralization Solution are from Promega Corporation.

Appendix Protocol 4 Minimal media**Minimal media:**

Autoclave	1	2.5 ml 20% glucose
		450 ml water
		0.5 g NH_4Cl
Filter sterilize		50 ml 10X M9 salts
		1 ml 1M MgSO_4
		50 μl 1M CaCl_2
		0.5 ml 50mg/ml carbenicillin

Combine and make 5 ml aliquots.

Maltose plates:

Autoclave	425 ml water
	7.5 g Agar
	0.5 g NH ₄ Cl
Filter sterilize	50 ml 10X M9 salts
	1 ml 1 M MgSO ₄
	10 ml 20% maltose
	50 µl 1 M CaCl ₂

Combine, add 0.5 ml 50mg/ml carbenicillin after some cooling, and pour plates.